# Des spectres MS/MS à l'identification des protéines - Interprétation des données issues de l'analyse d'un mélange de protéines d'un organisme non séquencé

Freddy Cliquet

**UNIVERSITÉ DE NANTES**

**UFR SCIENCES ET TECHNIQUES**

**ÉCOLE DOCTORALE**

**SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE MATHÉMATIQUES**

2011

**Thèse de Doctorat de l'Université de Nantes**

Spécialité : INFORMATIQUE

*Présentée et soutenue publiquement par*

**Freddy CLIQUET**

*le 27 juin 2011*

*à l'UFR Sciences et Techniques, Université de Nantes*

# Des spectres MS/MS à l'identification des protéines - Interprétation des données issues de l'analyse d'un mélange de protéines d'un organisme non séquencé

Jury

| | | |
|---|---|---|
| Rapporteurs : | Mme. Christine GASPIN, Directrice de recherches | INRA Toulouse |
| | M. Jacques NICOLAS, Directeur de recherches | IRISA/INRIA Rennes |
| Examinateurs : | M. Rémi HOULGATTE, Directeur de recherches | INSERM Nantes |
| | Mme. Hélène ROGNIAUX, Ingénieur de recherches | INRA Nantes |

**Directeur de thèse : Guillaume FERTIN**
**Co-directeur de thèse : Dominique TESSIER**
Laboratoire : LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE.
2, rue de la Houssinière, BP 92 208 – 44 322 Nantes, CEDEX 3.

N° ED 0366-...

# Des spectres MS/MS à l'identification des protéines - Interprétation des données issues de l'analyse d'un mélange de protéines d'un organisme non séquencé

---

## *From MS/MS spectra to protein identification - interpretation of data from shotgun protein analysis in an unsequenced organism*

**Freddy Cliquet**

⋈

*favet neptunus eunti*

---

**Université de Nantes**

Freddy CLIQUET

***Des spectres MS/MS à l'identification des protéines - Interprétation des données issues de l'analyse d'un mélange de protéines d'un organisme non séquencé***

x+134 p.

Ce document a été préparé avec LaTeX2e et la classe `these-LINA` version v. 1.30 de l'association de jeunes chercheurs en informatique LOGIN, Université de Nantes. La classe `these-LINA` est disponible à l'adresse :

`http://login.irin.sciences.univ-nantes.fr/`

*Impression : these.tex – 21/4/2011 – 9:17*

*Révision pour la classe : these-LINA.cls,v 1.30 2005/08/16 17:01:41 mancheron Exp*

## Résumé

La spectrométrie de masse est une technique utilisée en protéomique pour identifier des protéines inconnues dans un échantillon. Le spectromètre mesure la masse de fragments de la protéine et fournit ainsi des spectres expérimentaux qui sont des représentations, sous forme de séries de pics, de la présence de ces différents fragments. En étudiant ces spectres, nous espérons pouvoir identifier la protéine d'origine en la retrouvant dans une banque.
L'objectif de cette thèse est de proposer de nouvelles méthodes permettant d'étudier ces spectres. Cependant, ces méthodes doivent fonctionner sur des organismes non séquencés. Dans ce cas particulier, nous ne retrouverons pas exactement ces protéines dans la banque, mais uniquement des protéines qui y ressemblent.

Nous proposons tout d'abord un nouvel algorithme dit de comparaison de spectres : PacketSpectralAlignment. Cet algorithme permet de comparer des spectres expérimentaux à des spectres créés à partir des données contenues dans la banque, et ce, même en présence de modifications. Cette comparaison permet l'association de chacun des spectres à un peptide de cette banque. Ensuite, nous détaillerons différents prétraitements et filtrages permettant d'améliorer l'exploitation de notre nouvel algorithme. Tous ces éléments sont intégrés dans une plate-forme intitulé SIFpackets. Enfin, nous validons les résultats de PacketSpectralAlignment ainsi que de SIFpackets sur différents jeux de données réelles.



**Mots-clés :** Protéomique, Spectrométrie de masse, Comparaison de spectres, Identification de protéines, Modifications post-traductionnelles, Algorithmes, Programmation dynamique



## Abstract

Mass spectrometry is a general method used in proteomics to identify unknown proteins in a sample. The mass spectrometer measures the masses of several protein's fragments and provide spectra. A spectrum is a series of peaks that indicate the presence of those fragments. By studying these spectra, we aim at retrieving the analyzed protein in a reference databank. In this thesis, we propose a new method to study these spectra. However, our solution must be able to work on proteins coming from unsequenced species, which means that we can't find exactly the same proteins in the databank, only similar ones.

At first, we propose a new spectra comparison algorithm: PacketSpectralAlignment. This algorithm allows to compare experimental spectra produced by a mass spectrometer to spectra created from the reference databank data in presence of modifications. This comparison allows to associate to each spectrum, a peptide from the databank.
Then, we explain several preprocessing and filtering methods that enhance the results of our new algorithm. All of those methods are used in the SIFpackets framework.
Finally, we validate PacketSpectralAlignment and SIFpackets results using several experimental datasets.



**Keywords:** Proteomics, Mass spectrometry, Spectra comparison, Protein Identification, Post translational modifications, Algorithms, Dynamic programming

# Remerciements

Je tiens à remercier ici toutes les personnes m'ayant aidé tout au long de mon parcours.

En premier lieu, je remercie Dominique, Irena et Guillaume pour leur présence, leurs encouragements et conseils avisés qu'ils ont pu me dispenser tout au long de cette thèse. Merci d'avoir été si disponible, et ce malgré des emplois du temps pas toujours facile a concilier.

Je tiens à remercier Christine Gaspin et Jacques Nicolas d'avoir accepté d'être rapporteurs de ma thèse, ainsi que Hélène Rogniaux et Rémi Houlgatte pour avoir accepté d'être membre de mon jury de thèse. Un grand merci à vous.

Des remerciements pour les deux laboratoires m'ayant accueilli et plus particulièrement pour les équipes ComBi et Bioinfo dans leur intégralité. Des équipes pleines de personnes intéressantes et enrichissantes tant au niveau intellectuel que personnel. Ce fut un véritable plaisir de travailler près de ces personnes.

Ma famille et mes amis qui ont toujours été présent d'une manière ou d'une autre pour moi durant ces longs mois de thèse. Mes parents pour le soutien qu'ils m'ont apporté et sans lequel je n'aurais pu aller aussi loin. Mais aussi beaucoup de personnes qui se reconnaitrons ici, avec une pensé particulière pour Sylvie, Quentin, Alex, Izaskun et Christy pour leur soutien particulier.

Et enfin à Elsa, pour m'avoir supporté, écouté et avoir été présente auprès de moi tout au long de cette thèse.

À vous tous je dis merci.

# Sommaire

**1 Introduction** . . . . . **1**
**2 Notions de biologie et de protéomique** . . . . . **5**
2.1 Introduction . . . . . 5
2.2 Des gènes aux protéines . . . . . 6
2.3 Protéomique et spectrométrie de masse . . . . . 14
**3 L'identification de protéines en MS/MS - État de l'art et problématique** . . . . . **25**
3.1 Introduction . . . . . 25
3.2 L'interprétation *de novo* d'un spectre MS/MS . . . . . 25
3.3 L'identification par comparaison avec des protéines connues . . . . . 29
3.4 Comparaison des approches *de novo* et de PFF . . . . . 34
3.5 La problématique des modifications sans a priori . . . . . 34
**4 PacketSpectralAlignment, une nouvelle méthode de comparaison de spectres** . . . . . **45**
4.1 Introduction . . . . . 45
4.2 Notations . . . . . 45
4.3 Deux notions importantes : Symétrie et Paquets . . . . . 47
4.4 Modification des spectres . . . . . 48
4.5 Algorithme d'alignement de deux spectres . . . . . 53
**5 Jeux de données et critères d'évaluation** . . . . . **61**
5.1 Introduction . . . . . 61
5.2 Jeux de données . . . . . 61
5.3 Critères d'évaluation . . . . . 67
**6 SIFpackets : mettre PacketSpectralAlignment en situation réelle** . . . . . **69**
6.1 Introduction . . . . . 69
6.2 Amélioration de l'identification des peptides : paramétrage et prétraitements . . . 69
6.3 SIFpackets : une plate-forme complète associant spectres et peptides . . . . . 84
6.4 Remontée à la protéine . . . . . 96
**7 Conclusions et perspectives** . . . . . **101**
7.1 Conclusions . . . . . 101
7.2 Perspectives . . . . . 102

**Bibliographie** . . . . . **105**
**Liste des tableaux** . . . . . **113**
**Liste des figures** . . . . . **115**
**Liste des exemples** . . . . . **119**
**Table des matières** . . . . . **123**
**Index** . . . . . **125**

**A Évaluation du comportement de SpectralAlignment en présence de modifications . 131**

# CHAPITRE 1

# Introduction

Les protéines sont un des composés organiques essentiels à la vie. En effet, elles jouent un rôle central dans de nombreux processus biologiques, en accomplissant de très nombreuses fonctions, qui peuvent être aussi variées que la communication inter-cellulaire (e.g. les récepteurs d'hormones), la défense immunitaire (e.g. les immunoglobulines ou "anticorps"), le transport de certains éléments (e.g. l'hémoglobine qui transporte le dioxygène), le contrôle de la mobilité (e.g. l'actine et la myosine), mais aussi la construction et la maintenance des cellules (e.g. le collagène).

Dans beaucoup de projets de recherche, les protéines sont le centre d'intérêt des études effectuées ; cette branche de la recherche est appelée **protéomique**, par analogie avec le terme génomique.

En protéomique, la spectrométrie de masse est une méthode communément employée dans le but d'identifier des protéines. Dans ce processus, les protéines sont habituellement découpées en fragments, appelés peptides, qui seront ionisés et analysés par l'appareil. Le spectromètre de masse va pouvoir isoler ces peptides et mesurer un ensemble de masses les caractérisant. Ce processus va permettre d'obtenir, pour chaque peptide, un spectre qui se présente sous la forme d'une série de pics. Une fois les spectres obtenus, l'objectif est de les utiliser pour retrouver la composition de chaque peptide sous la forme d'une séquence de petites molécules appelées acides aminés. Ensuite, en recombinant ces peptides, il est généralement possible de retrouver, dans une banque de référence, la protéine analysée, qui est elle-même une longue séquence d'acides aminés.

Il existe deux grandes familles de méthodes permettant de retrouver, à partir des spectres, des séquences peptidiques : (i) l'approche dite *de novo*, qui va chercher à identifier chaque acide aminé sur des critères numériques (différences de masses), ou bien (ii) la comparaison de spectres qui va chercher à comparer les spectres à d'autres spectres créés de toutes pièces à partir des peptides contenus dans une banque de protéines : si deux spectres se ressemblent fortement, alors on peut raisonnablement penser qu'ils représentent le même peptide. Nous nous intéresserons plus particulièrement dans ce mémoire à la comparaison de spectres.

La démarche à base de comparaison de spectres présentée ci-dessus, bien que souvent très efficace, a des limites. En effet, pour être capable d'identifier un peptide en utilisant le contenu des banques, il faut être certain que celui-ci s'y trouve, c'est-à-dire que l'organisme dont sa protéine provient soit séquencé. Si les organismes les plus étudiés sont en très grande partie séquencés, certains ne le sont toujours pas. Cela se vérifie tout particulièrement en ce qui concerne les végétaux, centre d'intérêt particulier pour l'INRA de Nantes.

Lorsqu'il s'agit de travailler sur de tels organismes, il est cependant possible d'utiliser un orga-

nisme dit de référence, c'est-à-dire un organisme biologiquement proche en termes d'évolution. Ainsi, il est possible d'identifier des protéines d'un organisme à partir d'une banque constituée des protéines d'un autre organisme. Cependant, la tâche est plus complexe que pour un organisme séquencé. Dans ce cas, il est en effet possible que certaines protéines soient absentes, ou plus généralement qu'elles présentent des différences liées à l'évolution. Ces différences vont se traduire par des modifications de séquence protéique via des insertions, suppressions ou substitutions d'acides aminés.

Dans le cas d'organismes non séquencés, toute méthode visant à associer une séquence peptidique à un spectre devra donc être capable de prendre en compte ces différences. Cela complexifie fortement le problème et a intéressé plusieurs équipes ces dernières années [CMG+03, CB04, FP05, HGFA03, Mat07b, PDT00, SDW+05, TSYI03, TSF+05].

Nous allons aborder le problème de l'identification des protéines dans le cas d'organismes non séquencés, en proposant une nouvelle méthode de comparaison de spectres. Cette méthode sera, de plus, capable de tolérer un nombre important de modifications. Elle sera également capable de gérer le fait que l'on ne dispose d'aucune information a priori sur ces modifications. Enfin, notre méthode devra aussi être capable de localiser ces modifications dans le spectre, et ce dans le but d'aider les biologistes à mieux les comprendre et identifier.

## Organisation de ce mémoire

Nous introduisons tout d'abord dans le Chapitre 2 les notions importantes de biologie relatives aux protéines, puis nous présentons les principales techniques de la protéomique. Nous nous attardons ensuite plus longuement sur la spectrométrie de masse et les principes qu'elle met en oeuvre, entre autres dans l'objectif d'identifier des protéines. Nous y décrivons également la forme des données produites par les spectromètres de masse, ainsi qu'un protocole général pour les traiter.

Dans le Chapitre 3, nous présentons les différentes familles de méthodes existantes permettant l'identification de protéines à partir de spectres MS/MS. Nous y détaillons aussi bien les approches de type *de novo* que celles de type comparaison de spectres. Nous comparons ces différentes approches et discutons de leurs différentes faiblesses et points forts. Puis nous introduisons notre problématique de manière plus précise : l'identification des protéines lorsqu'elles ne proviennent pas d'organismes séquencés. Nous discutons finalement des difficultés supplémentaires que ce problème pose par rapport à l'approche classique utilisée sur les organismes séquencés.

Dans le Chapitre 4, nous introduisons l'algorithme de comparaison de spectres MS/MS que nous avons développé : PacketSpectralAlignment [CFRT09]. Il s'agit d'un algorithme s'appuyant sur la programmation dynamique et qui est capable de prendre en compte des modifications sans a priori.

Nous proposons ensuite, dans le Chapitre 5, différents jeux de données expérimentaux, ainsi que des critères d'évaluation qui vont nous servir à la fois à paramétrer notre méthode, mais aussi à nous comparer à des méthodes concurrentes.

Dans le Chapitre 6, nous décrivons et attribuons une valeur à chacun des paramètres nécessaires au bon fonctionnement de notre méthode. PacketSpectralAlignment est ensuite associé à des prétraitements visant à améliorer ses performances dans une plate-forme appelée SIFpackets [CFRT10]. Nous détaillons le fonctionnement de cette plate-forme et évaluons le gain de performances qu'elle apporte sur différents jeux de données expérimentales.

Enfin, nous concluons en rappelant les principaux résultats obtenus et en proposant de nouvelles perspectives de recherche afin d'améliorer notre solution d'identification de protéines dans le cadre d'organismes non séquencés.

CHAPITRE 2

# Notions de biologie et de protéomique

## 2.1 Introduction

Dans ce premier chapitre, nous introduirons un ensemble de notions indispensables à la compréhension de nos travaux. Nous rappellerons tout d'abord quelques notions de base sur le rôle des **protéines** dans les organismes vivants, et le mécanisme de leur synthèse. Nous insisterons sur la structure des protéines avec la description de leurs constituants de base, les acides aminés.

Contrairement au génome, qui reste constant dans les cellules d'un même organisme et au cours de la vie d'une cellule, le protéome, qui correspond à l'ensemble des protéines d'un organisme, subit des modifications permanentes en réponse aux différentes conditions environnementales. Un organisme possède donc une très grande diversité de protéomes que les approches protéomiques vont permettre de caractériser.

Le travail de cette thèse est centré sur l'identification de protéines d'organismes non séquencés, c'est-à-dire d'organismes dont les protéines ne sont pas encore référencées dans les banques de protéines. L'identification de ces protéines inconnues s'appuie sur les informations disponibles sur des protéines "proches", informations stockées dans des banques de données dont la plus généraliste que nous présenterons est UniProt. Nous définirons alors le terme de **modification** qui permet de distinguer deux protéines proches ou différentes formes d'une même protéine.

L'étude systématique des protéines a pu être envisagée grâce au développement de la **spectrométrie de masse**. Nous détaillerons les étapes d'un protocole d'analyse dans les approches de protéomique classique : séparation et hydrolyse des protéines, analyse de l'échantillon par spectrométrie de masse, interprétation des résultats pour l'identification des protéines contenues dans l'échantillon. Nous insisterons sur les contextes d'utilisation de la spectrométrie de masse en mode MS ou en mode MS/MS. Les travaux réalisés dans le cadre de cette thèse permettent d'interpréter des données obtenues en mode MS/MS.

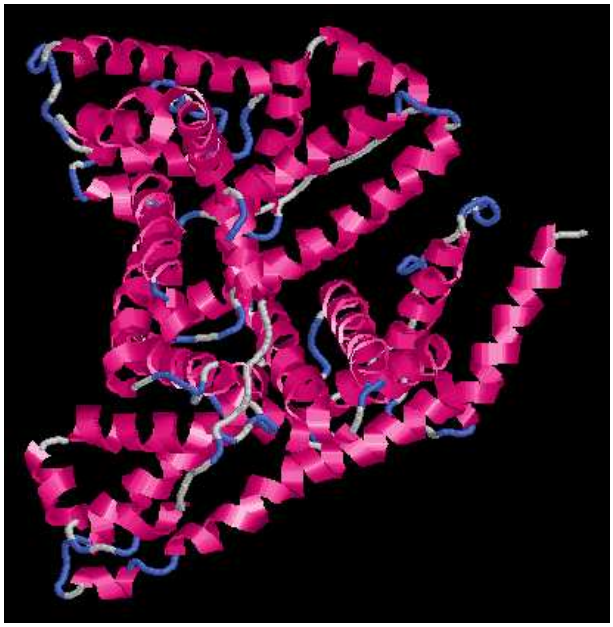

Figure 2.1 – Représentation 3D d'une protéine d'albumine de serum bovin (BSA). *Source : SWISS-MODEL,* http://swissmodel.expasy.org

## 2.2 Des gènes aux protéines

### 2.2.1 La cellule et ses protéines

La cellule est l'unité de base de tout organisme vivant. Elle coordonne une myriade de réactions biochimiques pour produire de l'énergie, synthétiser de nouveaux composants à partir de molécules organiques, répondre aux stimuli de l'environnement, maintenir et réparer les dommages causés à ses structures, grossir et se reproduire. Le matériel génétique, constitué d'un ensemble de **gènes** appelé **génome**, commande et programme la structure et le fonctionnement de toute cellule. Les gènes sont localisés dans les chromosomes et liés les uns aux autres de manière linéaire. Pour être utilisée dans la cellule, une partie de l'information génétique doit être décodée et transformée en **protéines**. En effet, ce sont ces macromolécules qui exécutent les principales fonctions cellulaires et qui assurent la construction et la maintenance de l'architecture de la cellule. Les protéines sont elle-mêmes composées par l'enchaînement de molécules plus simples : **les acides aminés**. Ceux-ci sont placés suivant un ordre précis qui caractérise la protéine, et que l'on appelle **la structure primaire** de la protéine.

Au sein de la cellule, les protéines ne se présentent pas sous une forme linéaire : elles se replient pour former une structure tridimensionnelle qui leur permet d'assurer correctement leurs fonctions biochimiques. La Figure 2.1 est une représentation de la structure tridimensionnelle d'une protéine d'albumine de sérum bovin.

#### 2.2.1.1 Synthèse d'une protéine

La synthèse d'une protéine se fait en deux étapes. Dans un premier temps, la séquence d'Acide DésoxyriboNucléique (ADN) codant le gène associé à la protéine est transcrite en Acide RiboNucléique messager (ARNm). Dans un second temps, l'ARNm est traduit en protéine.

**Transcription d'un gène.** Les gènes sont des fractions de chromosome qui codent des protéines et ont une nature chimique précise : ils sont composés d'acide désoxyribonucléique, ou ADN. Sur chaque désoxyribose est attachée une base azotée, formant ainsi ce que l'on nomme un **nucléotide**. Toute molécule d'ADN est constituée d'un enchaînement réalisé à partir de 4 nucléotides différents : A, G, C et T.

Ces gènes, ou séquences d'ADN, vont être transcrits en molécules d'ARNm dans le noyau de la cellule pour les organismes eucaryotes (organisme mono ou multicellulaire dont les cellules comportent un noyau) ou dans le cytoplasme pour les organismes procaryotes (organisme monocellulaire dans lequel la cellule ne comporte pas de noyau).

Cette transcription est une sorte de "copie" de l'ADN sous une forme légèrement différente, l'ARNm, qui va servir d'intermédiaire avant l'étape de traduction. Le but de cette copie est de préserver l'ADN (en le conservant dans le noyau pour les organismes eucaryotes), et d'augmenter la vitesse de production des protéines. L'étape de transcription est présentée à la Figure 2.2.

Dans le cas des eucaryotes, la transcription est complétée immédiatement par l'**épissage**. Durant cette étape, l'ARNm va être découpé et ligaturé dans le but d'en supprimer certaines régions. Les régions conservées sont appelées exons, tandis que les régions éliminées sont appelées introns. L'ARNm après l'épissage est qualifié d'**ARNm mature**.

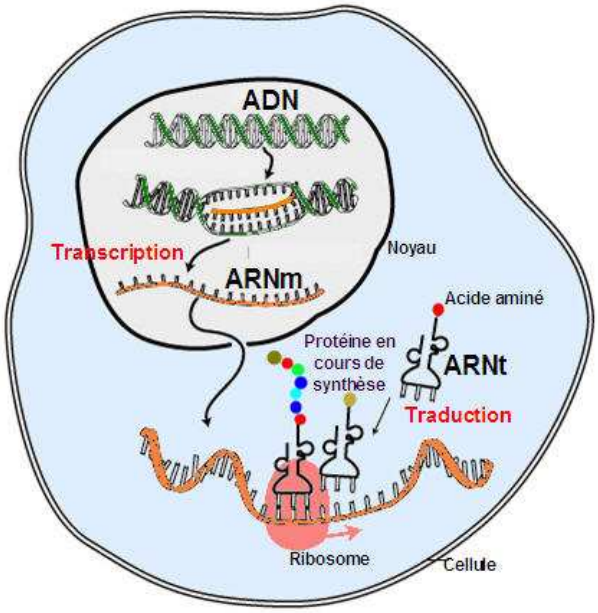


Figure 2.2 – Principe de la transcription de l'ADN en ARNm puis de la traduction de ce dernier en protéine. *Source : Journal du Net,* http://www.journaldunet.com/science/biologie/dossiers/06/0609-adn/adn2/transcription-traduction.jpg

**Traduction de l'ARN messager en protéines.** L'ARNm est ensuite traduit dans le cytoplasme, dans le but de produire la protéine. Les nucléotides de l'ARNm sont lus par triplets, un triplet étant appelé codon. Un codon va être traduit en un acide aminé ou en une instruction. Parmi les instructions importantes, nous pouvons mentionner le codon dit *initiateur*, qui va indiquer le début de la traduction, ou les codons dits *codon-stop*, qui indiquent que la fin de la séquence codant la protéine est atteinte. Un même brin d'ARNm peut servir à coder plusieurs exemplaires d'une même protéine. La Figure 2.2 présente la traduction de l'ARNm en protéine.

#### 2.2.1.2 Acides aminés et structure des protéines

La protéine est une macromolécule elle-même composée par l'enchaînement de molécules plus simples : les **acides aminés**. C'est l'enchaînement des quatre nucléotides A, G, C et T dans la séquence d'un gène qui va déterminer l'enchaînement des acides aminés au niveau de la protéine.

Un acide aminé est une molécule organique comprenant un squelette carboné, un groupement amine (-NH2), un groupement carboxylique (-COOH) et une chaîne latérale. Il existe naturellement 20 acides aminés communs à l'ensemble des espèces, tous différentiables par leur chaîne latérale, présentés Table 2.1 (en *vert* apparaît la chaîne latérale, en *bleu* le groupement amine et en *rouge* le groupement carboxylique). Les acides aminés sont désignés par un code international, composé d'une lettre ou de trois lettres, défini par l'IUPAC (International Union of Pure and Applied Chemistry) et l'IUBMB (International Union of Biochemistry and Molecular Biology). Les symboles utilisés sont présentés dans le Tableau 2.2 accompagnés d'abréviations complémentaires utilisées pour nommer certaines ambiguïtés. L'IUPAC définit aussi le **dalton** (noté Da) comme l'unité de masse utilisée pour évaluer la masse des protéines et de leurs constituants. Le dalton est équivalent à u, l'*unité de masse des atomes unifiée*, ce qui correspond à $1/12$ de la masse d'un atome $^{12}C$ de carbone, on a donc 1 Da = 1 u $\approx 1.660537781(82)$ x $10^{-27}$ kg.

La chaîne latérale d'un acide aminé lui confère des propriétés physico-chimiques particulières. Ces propriétés peuvent être déclinées en 5 catégories :

- acide
- basique
- neutre
- polaire (ou hydrophile)
- apolaire (ou hydrophobe).

Les acides aminés se lient entre eux avec des liaisons covalentes (appelées **liens peptidiques**) entre un groupement amine et un groupement carboxylique. Une chaîne d'acides aminés peut porter différents noms. Il est généralement admis qu'une chaîne de très petite taille (jusqu'à 5 acides aminés) sera nommée **tag**, qu'une chaîne comportant moins de 50 acides aminés sera nommée **peptide** et qu'une chaîne plus grande sera nommée **polypeptide**. Une protéine est quant à elle composée d'un ou plusieurs polypeptides. Le nombre d'acides aminés d'une protéine est très variable et peut aller de moins de cent jusqu'à plusieurs milliers.

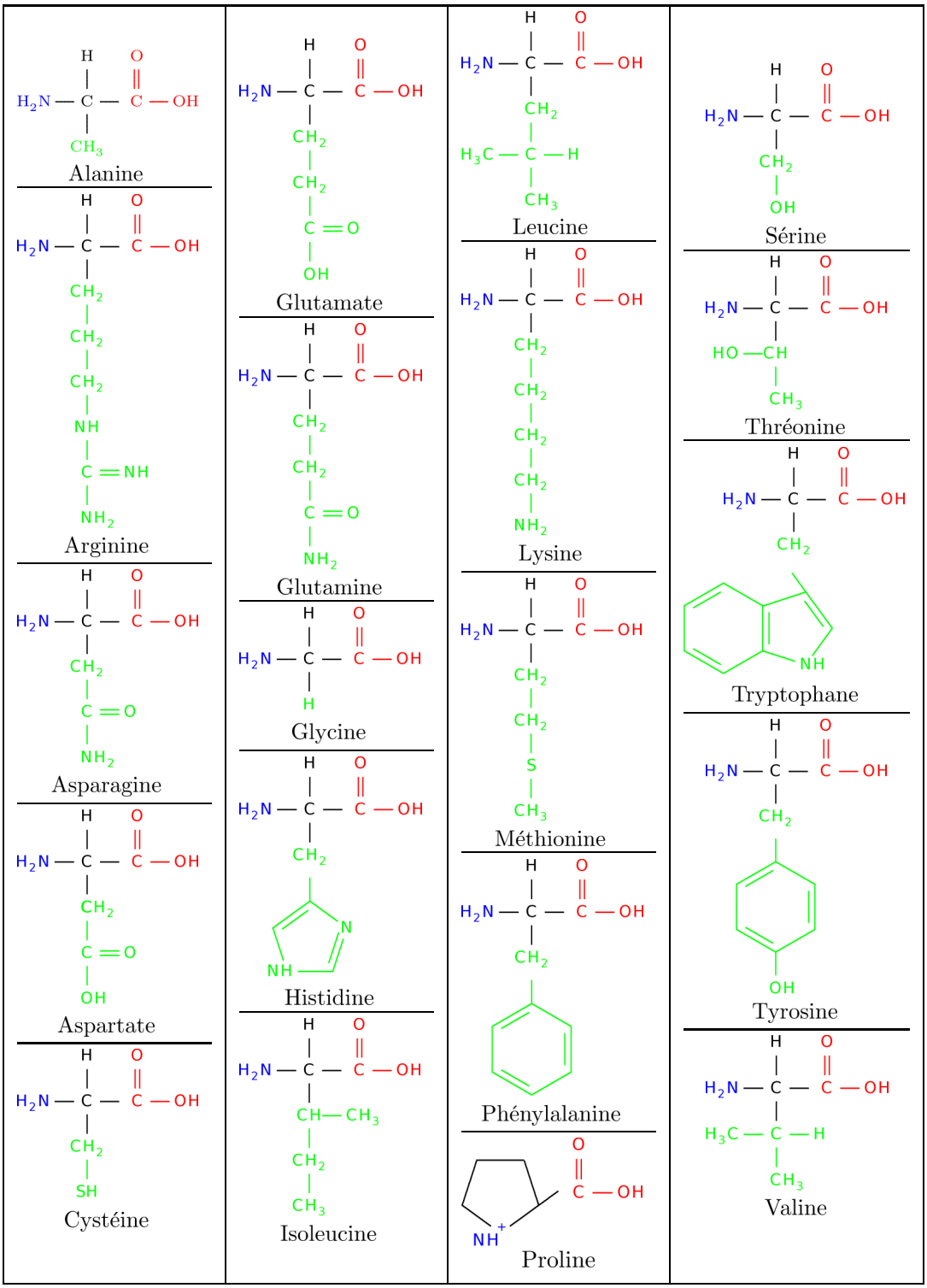


Table 2.1 – Dénomination et représentation de la structure des 20 acides aminés.

| Acide aminé | Abréviation en 3 lettres | Abréviation en 1 lettre | Masse mono-isotopique (Da) |
|---|---|---|---|
| Alanine | Ala | A | $71,03712$ |
| Arginine | Arg | R | $156,10112$ |
| Asparagine | Asn | N | $114,04293$ |
| Aspartate | Asp | D | $115,02695$ |
| Cystéine | Cys | C | $103,00919$ |
| Glutamate | Glu | E | $129,0426$ |
| Glutamine | Gln | Q | $128,05858$ |
| Glycine | Gly | G | $57,02147$ |
| Histidine | His | H | $137,05891$ |
| Isoleucine | Ile | I | $113,08407$ |
| Leucine | Leu | L | $113,08407$ |
| Lysine | Lys | K | $128,09497$ |
| Méthionine | Met | M | $131,04049$ |
| Phénylalanine | Phe | F | $147,06842$ |
| Proline | Pro | P | $97,05277$ |
| Sérine | Ser | S | $87,03203$ |
| Thréonine | Thr | T | $101,04768$ |
| Tryptophane | Trp | W | $186,07932$ |
| Tyrosine | Tyr | Y | $163,06333$ |
| Valine | Val | V | $99,06842$ |
| Asparagine ou Aspartate | Asx | B | $114,5 \pm 0.5$ |
| Glutamate ou Glutamine | Glx | Z | $128,5 \pm 0.5$ |
| Leucine ou Isoleucine | Xle | J | $113,08407$ |
| Acide aminé inconnu | Xaa | X | $NA$ |

Table 2.2 – Abréviations usuelles des 20 acides aminés accompagnées d'abréviations complémentaires utilisées pour les ambiguïtés. Les masses mono-isotopiques des différents acides aminés, c'est-à-dire les masses des isotopes les plus stables, sont également indiquées.

### 2.2.2 Modifications

#### 2.2.2.1 Mutations

Les mutations provoquent des variations dans la séquence d'acides aminés composant une protéine. Elles prennent leur source dans une variation de l'ADN, qui implique des changements lors de la synthèse de la protéine.

Il est possible d'observer ces mutations à deux échelles différentes :

- au niveau des individus, avec la variabilité intra-espèce,
- ou au niveau des espèces, avec l'évolution inter-espèces et l'adaptation des espèces à leur environnement.

Toutes ces mutations se traduisent, au niveau de la protéine, sous la forme de :

- Substitution : remplacement d'un acide aminé par un autre dans une séquence (Exemple 2.1).

– Insertion : ajout d'un ou plusieurs acides aminés consécutifs dans une séquence (Exemple 2.2).
– Suppression : suppression d'un ou plusieurs acides aminés consécutifs dans une séquence (Exemple 2.3).

Ces types de mutations sont visibles dans l'Exemple 2.4, où 2 protéines sont comparées. La première provient du maïs, la seconde du riz, deux céréales appartenant à la famille des *Poaceae*.

**Exemple 2.1 Substitution d'un acide aminé dans un peptide.**
*MCEEEDSTALVCDNGSGLCK* $\xrightarrow{substitution}$ *MCEEEDSTGLVCDNGSGLCK*

**Exemple 2.2 Insertion d'un acide aminé dans un peptide.**
*MCEEEDST A LVCDNGSGLCK* $\xrightarrow{insertion}$ *MCEEEDSTAQLVCDNGSGLCK*

**Exemple 2.3 Suppression d'un acide aminé dans un peptide.**
*MCEEEDSTALVCDNGSGLCK* $\xrightarrow{suppression}$ *MCEEEDS T LVCDNGSGLCK*

#### 2.2.2.2 Modifications Post-Traductionnelles

Une fois traduite, une protéine peut subir une maturation post-traductionnelle, c'est-à-dire une série de modifications biochimiques pouvant la modifier profondément. La protéine finale peut ainsi être très différente de la molécule directement codée par le gène. Il existe de très nombreuses variantes de modifications post-traductionnelles, comme celles modifiant la structure de la protéine (e.g. ponts disulfures). Cependant, les plus communes sont les ajouts / suppressions de composés chimiques sur les acides aminés. Ces modifications chimiques contribuent à la régulation de l'activité des protéines ainsi qu'à leur localisation dans les compartiments cellulaires.

Nous pouvons citer comme exemple de modification la *méthylation* qui consiste en l'ajout d'un groupement méthyle sur un acide aminé (généralement sur une lysine ou une arginine), ce qui a pour conséquence d'augmenter sa masse de 14 daltons, ou encore la *phosphorylation* qui est l'ajout d'un groupe phosphate à un acide aminé (généralement une sérine, thréonine, tyrosine ou histidine), ce qui a pour conséquence d'augmenter sa masse de 80 daltons. D'autres modifications peuvent aussi réduire la masse d'un acide aminé, par exemple la déshydratation qui réduira sa masse de 18 daltons.

**Exemple 2.4 Comparaison de protéines.** *Comparaison de la protéine* ***Granule-bound starch synthase 1*** *du Maïs (`SSG1_MAIZE`) avec celle du Riz (`LOC_Os06g04200.1`) , qui lui est génétiquement proche. L'alignement des deux protéines a été réalisé à l'aide de `ClustalW2` [LBB+07]*

– *un '.' sur la ligne du Riz signifie qu'il y a conservation de l'acide aminé (e.g. □)*
– *un acide aminé en minuscule sur la ligne du Riz signifie qu'il y a substitution (e.g. □)*
– *un '-' sur la ligne du Maïs signifie qu'il y a suppression d'un acide aminé dans le Maïs (e.g. □)*
– *un '-' sur la ligne du Riz signifie qu'il y a insertion d'un acide aminé dans le Maïs (e.g. □)*

```
Maïs  MAALATSQLVATRAGLGVPD--ASTFRRGAAQGLRG-ARASAAADTLSM   46
Riz   .s..t....atsat.f.ia.rsap.sll.hgf...kprsp.ggd.ts..v   50

Maïs  RTSARAAPRHQQQARRGGR-FPSLVVCAS-AGMNVVFVGAEMAPWSKTGG   94
Riz   t.....t.kq.rsvq..s.r...v..y.tg.....................  100

Maïs  LGDVLGGLPPAMAANGHRVMVVSPRYDQYKDAWDTSVVSEIKMGDGYETV  144
Riz   ......................i................a...va.r..r.  150

Maïs  RFFHCYKRGVDRVFVDHPLFLERVWGKTEEKIYGPVAGTDYRDNQLRFSL  194
Riz   .............i...s...k.....g......dt.v..k...m....  200

Maïs  LCQAALEAPRILSLNNNPYFSGPYGEDVVFVCNDWHTGPLSCYLKSNYQS  244
Riz   ............n.......k.t.................as...n...p  250

Maïs  HGIYRDAKTAFCIHNISYQGRFAFSDYPELNLPERFKSSFDFIDGYEKPV  294
Riz   n....n..v..............e.......s...r.........dt..  300

Maïs  EGRKINWMKAGILEADRVLTVSPYYAEELISGIARGCELDNIMRLTGITG  344
Riz   ..................................................  350

Maïs  IVNGMDVSEWDPSRDKYIAVKYDVSTAVEAKALNKEALQAEVGLPVDRNI  394
Riz   ............k....ta...at..i.............a......k.  400

Maïs  PLVAFIGRLEEQKGPDVMAAAIPQLMEMVEDVQIVLLGTGKKKFERMLMS  444
Riz   ..i.....................e..q--.................kl.k.  448

Maïs  AEEKFPGKVRAVVKFNAALAHHIMAGADVLAVTSRFEPCGLIQLQGMRYG  494
Riz   m...y............p...l.........p.................  498

Maïs  TPCACASTGGLVDTIIEGKTGFHMGRLSVDCNVVEPADVKKVATTLQRAI  544
Riz   .............v................k....s......a..k...  548

Maïs  KVVGTPAYEEMVRNCMIQDLSWKGPAKNWENVLLSLGVAGGEPGVEGEEI  594
Riz   ...............n.................g.....sa..i..d..  598

Maïs  APLAKENVAAP   605
Riz   ...........   609
```

Dans la suite de ce document, lorsque nous parlerons de **modifications**, ce terme inclura à la fois les mutations (substitutions, insertions et suppressions) et les modifications post-traductionnelles qui induisent un ajout ou une suppression d'un composé chimique.

### 2.2.3 Banques de protéines

Les connaissances accumulées sur les protéines sont répertoriées dans des banques de données. L'information la plus importante est certainement les séquences de protéines connues, obtenues via le séquençage de différents organismes. L'évolution des techniques, et le grand nombre de projets de séquençage actuel, impliquent une production toujours plus rapide de données.

En 2004, un important projet collaboratif a regroupé différentes banques de protéines existantes : the Universal Protein knowledgebase (UniProt) [ABW+04]. UniProt est structurée en deux parties : UniProt/SwissProt qui garantit une validation des données par des experts, et Uniprot/TrEMBL qui contient les traductions automatiques des séquences génomiques sans expertise complémentaire associée.

La quantité de données présentes dans ces bases augmente de manière exponentielle. Nous pouvons d'ailleurs voir sur la Figure 2.3 que le cap des 10 millions d'entrées a été franchi en 2010 pour Uniprot/TrEMBL. A titre de comparaison, UniProt/SwissProt contenait environ 500 000 entrées en 2010 (*selon l'European Bioinformatics Institute,* voir http://www.expasy.org/sprot/relnotes/relstat.html). Cela montre qu'il y a donc encore énormément de données non validées dans les banques.

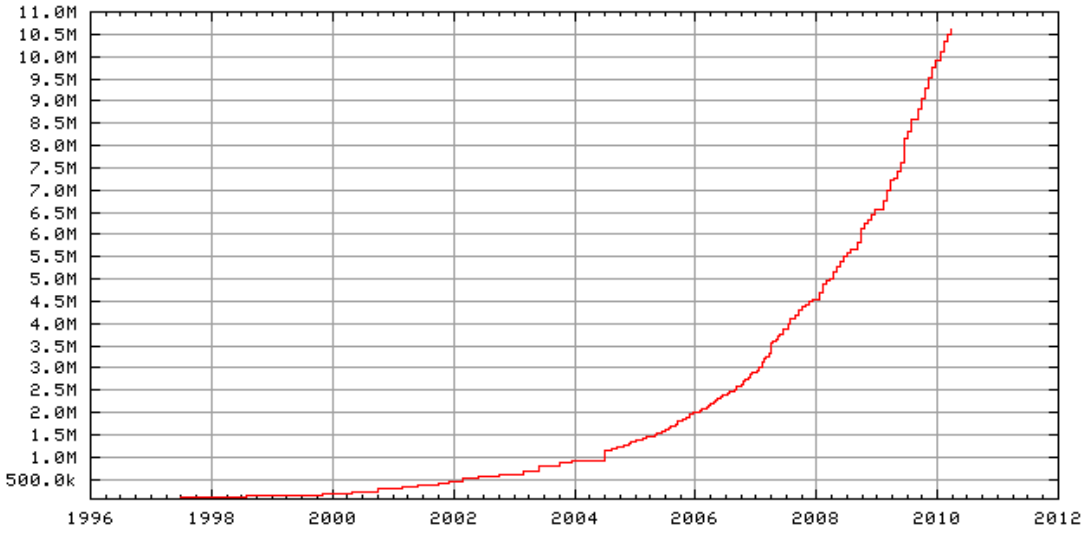


Figure 2.3 – Évolution du nombre d'entrées dans UniProt/TrEMBL au cours des 15 dernières années. *Source : European Bioinformatics Institute (EBI),* http://www.ebi.ac.uk/uniprot/TrEMBLstats/

La séquence d'une protéine peut être extraite des banques au format FASTA, qui donne pour chacune des protéines un identifiant accompagné d'un nom, suivi de la séquence codée en utilisant le code à une lettre présenté dans la Table 2.2. La séquence au format FASTA de la protéine Actin, aortic smooth muscle - Bos taurus (Bovine) est donnée dans l'Exemple 2.5.

**Exemple 2.5 Séquence de la protéine ACTA_BOVIN au format FASTA.**

```
>P62739|ACTA_BOVIN Actin, aortic smooth muscle - Bos taurus (Bovine).
MCEEEDSTALVCDNGSGLCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEA
QSKRGILTLKYPIEHGIITNWDDMEKIWHHSFYNELRVAPEEHPTLLTEAPLNPKANREK
MTQIMFETFNVPAMYVAIQAVLSLYASGRTTGIVLDSGDGVTHNVPIYEGYALPHAIMRL
DLAGRDLTDYLMKILTERGYSFVTTAEREIVRDIKEKLCYVALDFENEMATAASSSSLEK
SYELPDGQVITIGNERFRCPETLFQPSFIGMESAGIHETTYNSIMKCDIDIRKDLYANNV
LSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWIS
KQEYDEAGPSIVHRKCF
```

## 2.3 Protéomique et spectrométrie de masse

La protéomique est la science qui étudie les protéomes, c'est-à-dire l'ensemble des protéines d'un organisme. Les protéines sont étudiées depuis longtemps, nous pouvons par exemple citer Anselme Payen et Jean-François Persoz qui ont isolé la première enzyme, la diastase, en 1833 [PP33]. Mais les termes protéome et protéomique ne sont apparus que dans le milieu des années 1990, par analogie aux termes génome et génomique. Cette science ne s'est réellement développée que ces vingt dernières années.

La protéomique a des thématiques très variées pour lesquelles les méthodologies mises en oeuvre sont différentes les unes des autres. Parmi ces thématiques, nous pouvons citer :
- l'identification des protéines dans un échantillon donné,
- la localisation des protéines dans un organisme,
- la quantification des protéines en fonction du temps, de leur environnement ou encore de leur état,
- la caractérisation des modifications post-traductionnelles,
- la caractérisation des interactions des protéines (par exemple avec d'autres protéines),
- la caractérisation de la structure spatiale des protéines.

Dans la suite de notre travail, nous nous intéressons à l'identification de protéines ainsi qu'à l'étude des modifications post-traductionnelles. C'est pourquoi nous ne détaillerons que les méthodes couramment utilisées dans ces domaines.

### 2.3.1 Introduction à l'identification de protéines

Une analyse de protéomique classique visant à identifier des protéines peut généralement se décomposer en plusieurs étapes importantes, présentées Figure 2.4.

Tout d'abord, des protéines sont extraites d'un échantillon d'intérêt, avant d'être séparées à l'aide d'une technique dite de "séparation des données". Durant cette étape, les protéines sont hydrolysées, c'est-à-dire découpées en fragments peptidiques.

Dans un second temps, les peptides ainsi formés vont être analysés via la spectrométrie de masse (MS) afin d'obtenir leur masse avec une très grande précision, ou via la spectrométrie de masse en tandem (MS/MS) pour obtenir des informations sur leur composition en acides aminés.

Enfin, pour terminer, les résultats de ces analyses de spectrométrie de masse vont pouvoir être étudiés et confrontés à des informations contenues dans des banques de protéines, afin d'identifier la ou les protéine(s) analysée(s).

### 2.3.2 Séparation des protéines

Une analyse protéomique d'identification ou de quantification de protéines se base sur l'analyse d'un échantillon contenant un extrait de protéines. Cet échantillon peut contenir une très grande quantité de protéines différentes, dans des quantités variables. Pour en simplifier l'analyse, il est possible d'utiliser des stratégies dites de séparation des données. Ces techniques permettent de sélectionner les protéines d'intérêts de l'échantillon, ce qui améliore grandement les performances des procédures d'analyse. Les techniques de séparation peuvent s'échelonner

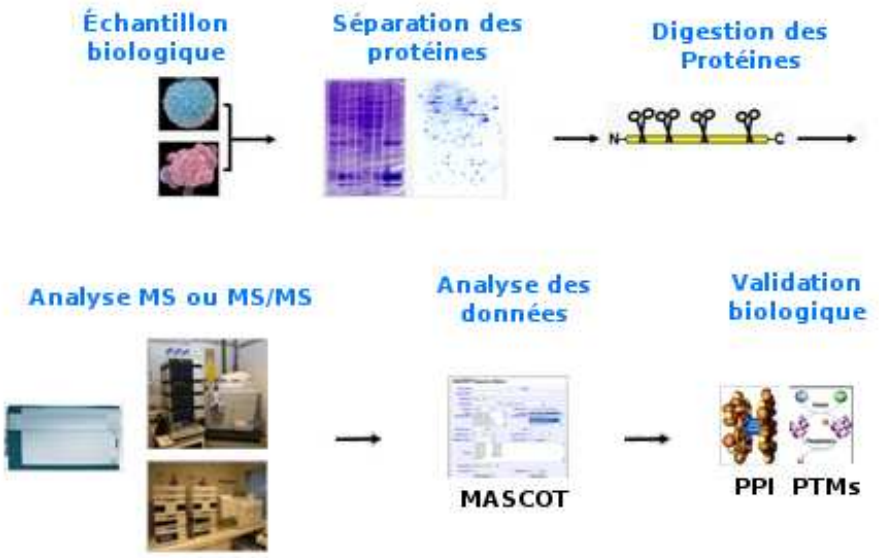


Figure 2.4 – Illustration de la démarche d'une analyse classique en protéomique. *Source : University of Oslo,* http://www.biotek.uio.no

en plusieurs étapes selon différentes caractéristiques physiques ou chimiques. Il est fait mention de ces méthodes sous les termes 2D, 3D, voire multidimensionnelles selon le nombre d'étapes de séparation ayant lieu. Les deux méthodes les plus couramment utilisées sont la chromatographie et l'électrophorèse.

#### 2.3.2.1 La Chromatographie

La chromatographie est une technique de séparation des données qui consiste à faire passer un mélange dissous dans une **phase mobile** (gazeuse ou liquide) au travers d'une **phase stationnaire** (gazeuse, solide ou liquide). Les différents éléments du mélange vont se déplacer à une vitesse particulière, dépendante de leurs propres caractéristiques ainsi que de celles de la phase mobile et de la phase stationnaire. La chromatographie peut-être principalement utilisée de deux façons : soit de manière **analytique** pour mesurer des proportions de composés dans un mélange, soit de manière **préparative** pour effectuer une purification en vue d'un usage futur. Il existe différents types de chromatographie qui sont définis en fonction de leur différentes phases. Celle qui est la plus communément couplée à la spectrométrie de masse est la chromatographie en phase liquide (appelée LC pour *Liquid Chromatography*).

**Chromatographie en Phase Liquide.** Cette chromatographie utilise, comme son nom l'indique, une phase mobile liquide. Cette phase mobile va circuler dans une colonne contenant la phase stationnaire qui peut être soit liquide, soit solide. La phase stationnaire doit avoir des interactions avec les éléments analysés. En faisant varier les conditions expérimentales, par exemple en changeant le pH de la solution, il est possible de libérer ou de retenir certains types d'éléments et ainsi de filtrer les éléments sortant de la colonne. Les interactions avec la phase

stationnaire peuvent être de différents types, comme :

– *la chromatographie en phase inversée* (RPLC pour *Reverse Phase Liquid Chromatography*), la plus communément employée, permet de séparer les éléments de la solution suivant leur hydrophobicité.

– *la chromatographie à exclusion de taille* (SEC pour *Size Exclusion Chromatography*) permet de séparer les éléments selon leur taille ou leur volume hydrodynamique, c'est-à-dire en fonction du volume occupé par l'élément quand il est dans une solution. Il est à noter que ce volume dépend à la fois de l'élément et de la solution utilisée.

#### 2.3.2.2 Électrophorèse

Cette technique a pour principe de séparer les particules d'une solution en utilisant leur charge électrique, complétée par d'autres caractéristiques comme la taille de chacune des particules. Les éléments à séparer, après avoir été ionisés, vont se déplacer sur un support sous l'effet d'un champ électrique. Dans un deuxième temps, il est possible de séparer les éléments de manière orthogonale sur le même support pour ainsi former une électrophorèse bi-dimensionnelle. Cette deuxième séparation peut se faire sur la base du temps de migration des différents éléments au travers du support qui est poreux, un petit élément migrant plus vite qu'un grand élément. Le support employé pour faire migrer les protéines est un gel, appelé **gel d'électrophorèse**.

Il est courant en protéomique d'utiliser des gels d'électrophorèse 2D. Dans un tel gel, les protéines sont séparées en différents petits amas appelés **spots**. Un spot représente des protéines ayant des caractéristiques similaires (charge électrique, taille). La Figure 2.5 présente un gel 2D. Lorsque la technique de séparation est optimale, un spot contient un nombre important de copies d'un nombre très limité de protéines.

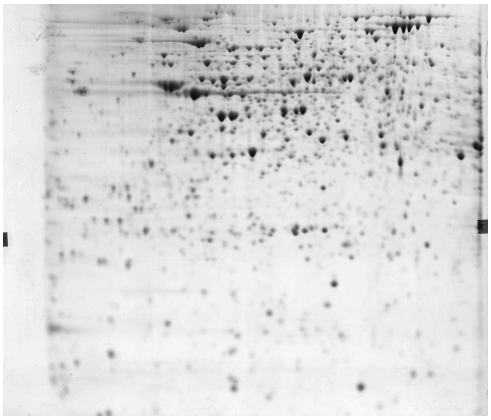

Figure 2.5 – Gel d'électrophorèse 2D. *Source : Institut National de Recherche en Agronomie (INRA),* http://pappso.inra.fr/2DE/application.html

### 2.3.3 Hydrolyse des protéines

L'étape d'hydrolyse des protéines permet de rompre des liaisons peptidiques dans une protéine, et ainsi d'obtenir des fragments plus petits : des peptides. Différents agents peuvent être utilisés pour causer cette hydrolyse. Chacun de ces agents a pour particularité de couper une séquence protéique après certains acides aminés spécifiques. Il est donc possible, en fonction de la composition en acides aminés des protéines d'intérêt, de choisir l'agent d'hydrolyse qui produira des peptides d'une taille moyenne intéressante, cette taille moyenne intéressante étant déterminée par la taille des peptides pouvant être analysés ensuite par le spectromètre de masse.

La majorité des agents employés sont des **enzymes**, la plus fréquente étant la **trypsine**, qui produit de nombreux peptides de taille exploitable pour la plupart des protéines généralement analysées. La trypsine coupe les liaisons peptidiques qui suivent les lysines et les arginines, sauf si elles sont immédiatement suivies d'une proline. Des agents chimiques, comme le bromure de cyanogène (qui coupe la protéine après les méthionines) peuvent aussi être employés.

### 2.3.4 Spectrométrie de masse

#### 2.3.4.1 Présentation de la technique d'analyse

La spectrométrie de masse est une technique d'analyse qui peut être utilisée pour identifier des molécules. Un **spectromètre de masse** est un appareil comportant une source d'ionisation qui permet d'ioniser les molécules, et un ou plusieurs analyseurs de masse qui vont fournir en sortie des informations sous la forme d'un spectre de masse. Le spectre va contenir des pics d'intensité variable, localisés à des positions (abscisses) représentant le ratio entre la masse $m$ et la charge $z$ des molécules ($m/z$). L'interprétation de l'information portée par ce spectre peut permettre d'identifier la ou les protéine(s) analysée(s). Plus un pic a une intensité importante (i.e. plus il est grand), plus nous avons une certitude forte qu'un élément avec le $m/z$ donné est présent dans l'échantillon analysé. A contrario, une faible intensité, ou la non présence d'un pic à une position donnée, ne garantit pas que l'élément soit absent de l'échantillon.

**Ionisation.** Le spectromètre de masse nécessite que chaque molécule soit chargée. Il faut donc une source d'ionisation pour vaporiser et ioniser les molécules. Il existe deux sortes de sources :

- *Ionisation par électronébuliseur* (appelée `ESI` pour *ElectroSpray Ionization*) : elle permet de disperser un liquide sous forme de gouttelettes chargées électriquement. La charge générée par la source `ESI` peut être simple ou multiple. Cette méthode, créée par John Bennet Fenn en 1989 [FMM+89], lui a valu le prix Nobel de Chimie en 2002.

- *La désorption-ionisation laser assistée par matrice* (appelée `MALDI` pour *Matrix-Assisted Laser Desorption/Ionization*) permet de vaporiser et ioniser des molécules enfermées dans une matrice à l'aide d'un laser. La matrice protège les molécules de la destruction au contact du faisceau laser. Les molécules ionisées par cette méthode ont la particularité d'être simplement chargées. Cette technique a tout d'abord été développée par Hillenkamp et Karas en 1985 [KBH85], puis améliorée par Tanaka en 1988 [TWI+88], ce qui valut à ce dernier le prix Nobel de Chimie en 2002.

Généralement, les sources de type `MALDI` sont associées à une séparation sur gel d'électrophorèse 2D, et les sources de type `ESI` à une chromatographie liquide. Le choix entre ces deux sources d'ionisation repose sur plusieurs critères tels que :

- le volume de l'échantillon : si très peu de produit est disponible, généralement la technique gel d'électrophorèse 2D, puis source `MALDI`, sera retenue.

- la taille des molécules (peptides) : les appareils présentent généralement une tolérance maximale en terme de $m/z$. Si les peptides sont très gros, il faut privilégier une source `ESI` qui, avec une charge plus importante qu'une source `MALDI`, va réduire le ratio $m/z$.

**Analyseur de Masse.** Les analyseurs de masse utilisent différents principes pour séparer les molécules étudiées en fonction de leur ratio $m/z$, où $m$ est la masse de la molécule et $z$ sa charge. Pour cela, plusieurs analyseurs existent et sont couramment utilisés :

- Le *Temps-de-Vol* (appelé `TOF` pour *Time-of-Flight*) mesure le temps mis par un ion accéléré par une tension pour parcourir une distance donnée. Ce temps de vol permet d'obtenir directement le ratio $m/z$.

- Le *Quadripôle* (`Q`) est constitué de quatre tiges métalliques parallèles et placées de manière symétrique. Les barres sont soumises à un potentiel électrique positif ou négatif (une barre sur deux) variable. Cela va faire osciller les ions traversant la zone comprise entre les barres. En fonction de l'amplitude de variation du potentiel, ainsi que de la fréquence de variation de ce potentiel appliqué aux barres, les ions ayant un certain ratio $m/z$ pourront soit traverser la zone, soit être déviés hors de celle-ci. Le fonctionnement d'un quadripôle est illustré à la Figure 2.6.

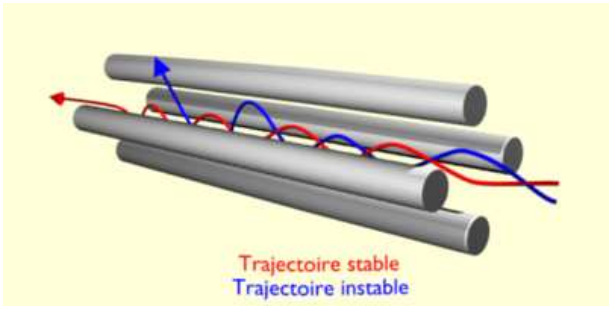


Figure 2.6 – Représentation d'un analyseur de masse type *Quadripôle*. En rouge figure la trajectoire d'un ion ayant un $m/z$ sélectionné pour traverser l'analyseur, en bleu un ion éjecté car son $m/z$ n'a pas été sélectionné. *Source : Wikipedia.*

- Le *Piège à Ions* (*Ion traps*) enferme plusieurs ions dans un espace réduit et est capable de les conserver à l'intérieur pendant un temps important. Il les conserve en faisant varier le potentiel électrique sur deux électrodes circulaires. La fréquence des oscillations de chacun des ions est dépendante de leur $m/z$. En changeant le potentiel électrique et la vitesse

de variation, il est possible d'éjecter des ions ayant un $m/z$ donné.

– L'analyseur à *Transformée de Fourier* (appelée FT pour *Fourier Transform*) est constitué d'une chambre entourée d'un champ magnétique très fort. Les ions injectés à l'intérieur vont tourner avec le champ magnétique à une fréquence dépendant de leur $m/z$. La variation du champ magnétique produit une variation dans la fréquence de rotation des ions, ce qui permet, en appliquant une transformée de Fourier, d'obtenir leur ratio $m/z$.

Un analyseur de masse peut être évalué sur de nombreux critères : sa vitesse de traitement, sa fenêtre de tolérance, sa résolution, sa précision et son prix. La fenêtre de tolérance est définie par un $m/z$ minimal et maximal. La résolution est la capacité à discriminer des ions ayant des $m/z$ différents. La précision évalue l'erreur entre le vrai $m/z$ et celui trouvé par l'appareil. Elle est généralement exprimée en *daltons* ou en *parties par millions* (*ppm*).

Chaque analyseur présente des avantages et des inconvénients. Le FT est certainement le plus précis et a une résolution très élevée, mais il est très cher. Les analyseurs de type TOF ou Q offrent une résolution moyenne, mais une excellente précision. Et enfin les pièges à ions, qui ont des caractéristiques plutôt moyennes, offrent l'avantage d'être très robustes, et peu chers. Le type d'analyseur utilisé dépend généralement de l'usage souhaité.

#### 2.3.4.2 Spectrométrie de masse MS

Dans le contexte de la spectrométrie MS, les peptides issus de la digestion des protéines via une enzyme (généralement la trypsine) sont analysés par l'appareil. Celui-ci va mesurer, pour l'ensemble de ces peptides, leur ratio $m/z$. L'appareil va donc produire un spectre de masse pour chacune des protéines analysées, chacun des spectres contenant des pics représentant les peptides de cette protéine. Cependant, un analyseur de masse peut manquer certains peptides pour des raisons telles que :

– une fenêtre de tolérance insuffisante (le peptide manqué est donc trop lourd ou trop léger)
– un problème lors de la séparation des peptides (*e.g.* un peptide très hydrophobe peut ne pas passer dans une LC)
– un mauvais découpage de la protéine en peptides (problème lié à l'enzyme utilisée)

Les appareils utilisés le plus fréquemment en spectrométrie de masse MS, sont les MALDI-TOF et les ESI-TOF. Par exemple, la Figure 2.7 présente un spectre MS obtenu sur un appareil de type ESI-TOF.

Pour être interprétable, un spectre MS nécessite généralement des prétraitements (ou *pre-processing*). Il faut calculer les **centroïdes des pics** (discrétisation du spectre brut), **éliminer le bruit**, ajuster le **calibrage** (décalage de tous les $m/z$ pour compenser un mauvais calibrage de l'appareil) et **dé-isotoper** les spectres (retirer les pics marquant des isotopes). Ces traitements sont généralement effectués par un logiciel propriétaire, fourni par le fabricant car il est spécifique à chaque type de spectromètre de masse.

La spectrométrie de masse MS est généralement utilisée pour analyser des échantillons simples, c'est-à-dire contenant une seule protéine, et présentant peu de modifications. Cette méthode s'applique plus facilement à des organismes très bien connus et annotés dans les banques.

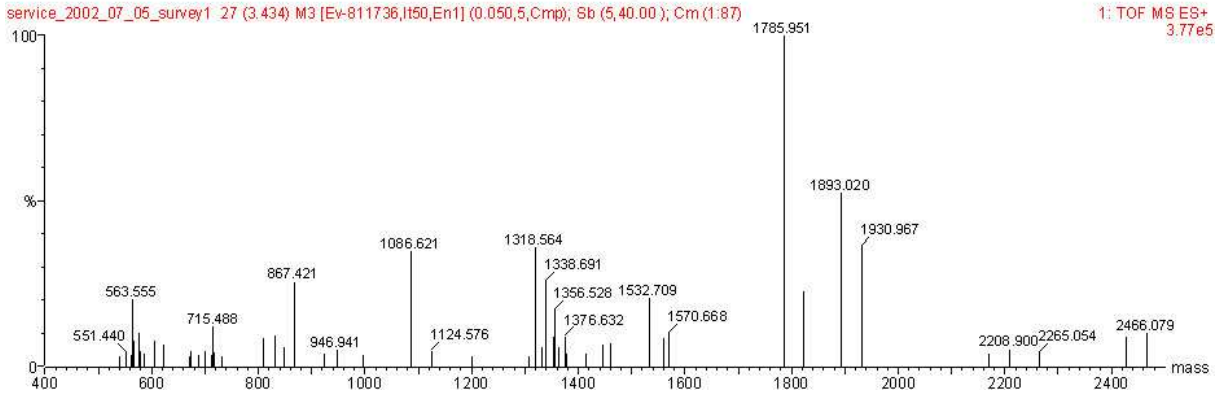


Figure 2.7 – Spectre MS obtenu après l'analyse d'une protéine avec un appareil de type ESI-TOF. Chaque pic possède une étiquette donnant sa masse (position en abscisse). L'ordonnée représente l'intensité d'un pic. *Source : Molbiotech.*

**Identification des protéines.** L'approche utilisée pour identifier les protéines en MS est généralement appelée *Peptide Mass Fingerprinting* (PMF). Elle a été développée en 1993 par plusieurs groupes indépendants [PHB93, HBS$^+$93, MHR93, JQCG93, YSGH93]. Elle consiste à comparer les spectres de masse fournis par l'appareil à des spectres théoriques. Ces derniers sont obtenus en simulant *in-silico* le comportement de l'enzyme utilisée expérimentalement sur une banque de protéines. Donc, pour chacune des protéines de la banque, un spectre théorique sera calculé. La comparaison consiste à évaluer la similarité entre un spectre expérimental et un spectre théorique. Par exemple, un critère de similarité peut être le nombre de pics en commun : plus le nombre de pics en commun est important, plus les spectres (et donc les protéines) seront supposés **similaires**. Les résultats sont classés par similarité, la meilleure protéine étant considérée comme l'identification correcte, si le nombre de pics en commun est suffisant.

**Quelques méthodes de PMF.** Le principe des méthodes PMF repose toujours sur les mêmes bases. La principale différence entre les méthodes existantes réside dans la fonction de score, c'est-à-dire la manière d'évaluer la similarité entre un spectre expérimental et un spectre théorique. Le décalage entre les pics, l'intensité, les pics manquants, le bruit, les modifications sont autant d'éléments pouvant être pris en compte dans le score. Nous pouvons donc classer les méthodes de PMF en fonction du score qu'elles utilisent :

– Le **nombre de masses partagées** entre le spectre expérimental et les masses théoriques des peptides d'une protéine est un indicateur fréquent. Cette méthode de score n'est pas très sensible au bruit au sein des spectres MS, mais elle n'est en revanche pas très discriminante. En effet, la taille de la banque de protéines influence très fortement la qualité

des identifications : une banque trop grande offre généralement des identifications ambiguës [GM01]. Des applications comme PepSea [MHR93], PeptIdent [WGB+99] ou encore PeptideSearch [MW94] utilisent ce système de score.

– Une **approche statistique ou probabiliste** sert aussi souvent de fonction de score. La plus connue est certainement MOWSE [PHB93], qui sera à la base du score de nombreuses autres méthodes telles que MS-Fit [CHS+95], ProFound [ZC00] ou encore le très utilisé Mascot [PPCC99]. Le score de MOWSE est issu de l'analyse statistique de la fréquence d'apparition des masses des peptides dans la banque de protéines OWL [BAA94]. Si Mascot a son score basé sur MOWSE, il intègre de nombreuses autres informations sur la précision de la correspondance entre le spectre théorique et le spectre expérimental, ainsi que sur la prise en compte de modifications. En revanche, aucun détail sur cette partie du score n'est divulgué par *Matrix Science Inc.*, la société éditrice de l'application.

– L'**apprentissage automatique** est employé dans un outil comme SmartIdent [GMG+99]. SmartIdent utilise un algorithme génétique afin d'optimiser un score heuristique dans le but de fournir une identification automatique de protéines. Pour cela, l'algorithme fait intervenir toutes sortes d'informations provenant à la fois des données expérimentales et théoriques. L'objectif principal de cet outil est de retirer toute intervention de l'expert, à quelque étape que ce soit dans le processus d'identification. Cette méthode prend en compte de nombreux éléments dans son score, ce qui lui permet une forte discrimination [GM01].

La spectrométrie de masse MS reste encore efficace pour identifier une protéine unique, ou des mélanges contenant peu de protéines. En revanche, cette méthode montre ses limites lorsqu'il s'agit d'identifier les protéines de mélanges plus complexes (aussi appelé "shotgun"), ou lorsque des modifications sont présentes dans les protéines analysées [HWS03, Mat07a]. Dans de tels cas, l'usage de la spectrométrie de masse en tandem sera à préconiser.

#### 2.3.4.3 La spectrométrie de masse en tandem ou MS/MS

La spectrométrie de masse en tandem (ou MS/MS) consiste à combiner deux analyseurs. Ce système va permettre d'augmenter la précision des analyses.

La Figure 2.8 illustre le fonctionnement d'un spectromètre de masse en mode MS/MS. Les ions sont analysés dans un premier analyseur, puis passent dans une chambre de collision où ils sont fragmentés. Les fragments sont ensuite envoyés dans un second analyseur de masse.

Les deux analyseurs ne sont pas nécessairement du même type. L'utilisation d'analyseurs de types différents permet de bénéficier d'avantages plus variés, ou de limiter les inconvénients d'un analyseur donné. Par exemple, le « Quadrupole Time-of-flight » (aussi appelé `QTOF`) est un appareil utilisant un analyseur de type quadripôle dans un premier temps, avant de fragmenter les ions et de les faire passer dans un analyseur de type Temps-de-Vol.

**Fragmentation.** La fragmentation en MS/MS va couper le peptide à différents endroits, et ainsi créer une série d'ions. Cette fragmentation a principalement lieu sur le squelette peptidique. Selon la notation introduite en 1984 par Roepstorff et Fohlman [RF84], les ions formés

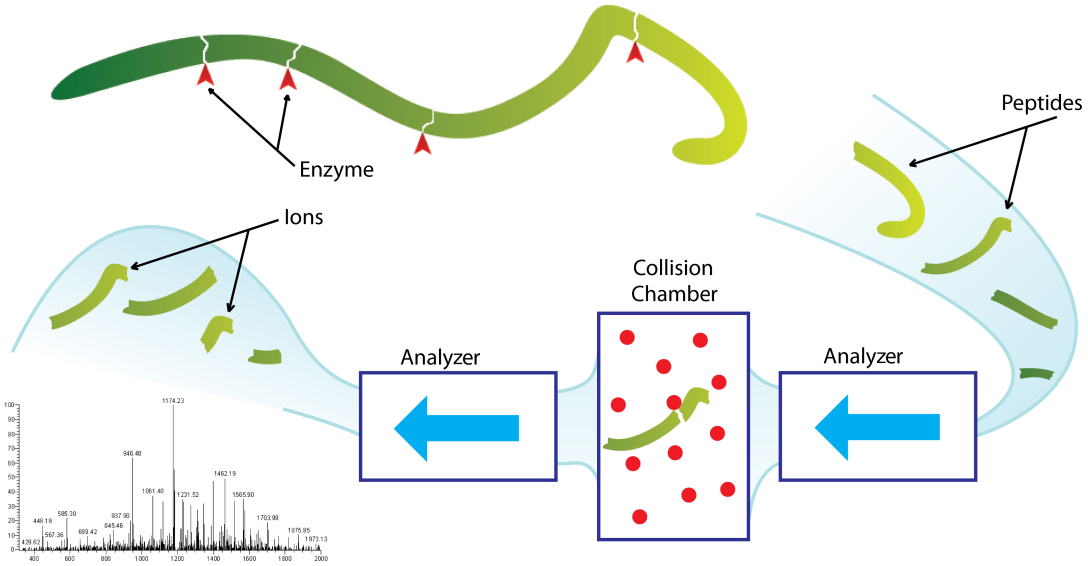


Figure 2.8 – Principe général d'un spectromètre de masse MS/MS.

quand la charge reste sur le fragment de gauche (nommé **N-terminal**) sont appelés $a$, $b$ et $c$ ; quand la charge reste sur le fragment de droite (nommé **C-terminal**), les ions formés sont appelés $x$, $y$ et $z$. La Figure 2.9 présente la position de la fragmentation pour ces différents ions. La détection de chacun de ces ions va se traduire par un pic dans le spectre MS/MS.

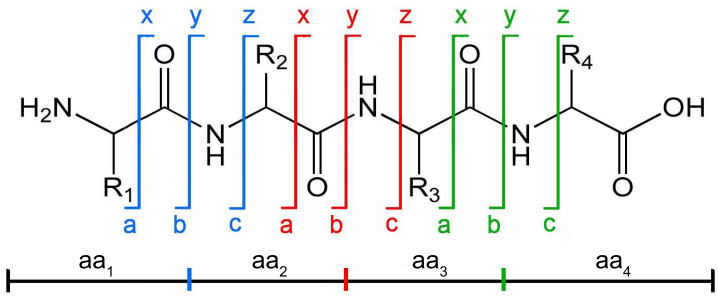


Figure 2.9 – Les différents sites de fragmentation à l'intérieur d'un peptide de quatre acides aminés ($AA_i$ avec $i \in [1;4]$). $R_i$ ($i \in [1;4]$) représente la chaîne latérale de $AA_i$.

**Spectre MS/MS.** Comme pour les spectres MS, les spectres MS/MS nécessitent des prétraitements pour être interprétables (calcul des centroïdes, élimination du bruit, calibrage, déisotopage). Tout comme dans les spectres MS, les pics des spectres MS/MS sont caractérisés par leur position ($m/z$ d'un fragment) et leur intensité. La comparaison s'arrête cependant là : un spectre MS/MS est plus complexe, et comporte potentiellement de très nombreux pics significatifs ayant des origines variées, par opposition au spectre MS dans lequel un pic significatif représente nécessairement le $m/z$ d'un peptide.

Le nombre de pics présents dans un spectre MS/MS est très variable et dépend de la longueur du peptide, mais aussi de la qualité de la fragmentation (certains acides aminés se fragmentent en effet mieux que d'autres), du type de spectromètre ou encore des pré-traitements appliqués. Les Figures 2.10, 2.11 et 2.12 présentent différents spectres expérimentaux de qualité variable provenant de l'analyse d'une protéine de Brachypodium sur un spectromètre MS/MS de type `QTOF`. Les spectres 1. et 2. de la Figure 2.10 montrent qu'un même peptide (SFRPLADHDVR dans ce cas) peut produire des spectres sensiblement différents. Dans la Figure 2.11, les spectres 1. et 2., représentant respectivement les peptides GGGDDNNNLQIACFEIR et EGDVIVAPAGTLMYLANTDGR, illustrent le fait qu'un spectre peut avoir des pics marquant très clairement les acides aminés dans une zone (ici le milieu des spectres) et des zones donnant pas ou peu d'informations (ici le début et la fin des spectres). Le spectre de la Figure 2.12, représentant le peptide QQQQQEGEEEGFIIR, illustre le fait qu'un spectre peut présenter une faible intensité pour une grande majorité ses pics, ce qui rend alors l'interprétation des pics difficile.

Le peptide ayant été fragmenté pour produire un spectre MS/MS est appelé **précurseur**. La masse du peptide précurseur est appelée **masse du précurseur**.

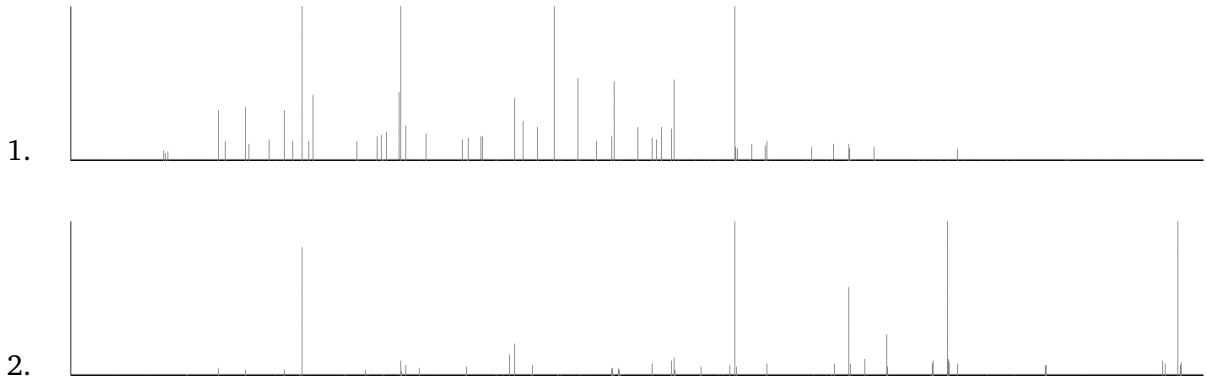


Figure 2.10 – Les deux spectres MS/MS présentés dans cette figure représentent un même peptide (SFRPLADHDVR), et ont tous deux été obtenus à partir de l'analyse d'une protéine de Brachypodium sur un spectromètre MS/MS de type `QTOF`. Ces deux spectres illustrent la variabilité du résultat de l'analyse d'un même peptide par un même spectromètre de masse.

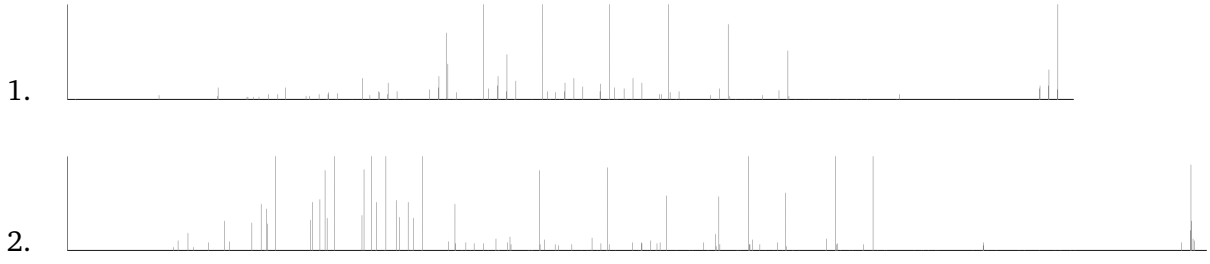


Figure 2.11 – Les spectres MS/MS 1. et 2. représentent respectivement les peptides GGGDDNNNLQIACFEIR et EGDVIVAPAGTLMYLANTDGR. Ces spectres sont issus de l'analyse d'une protéine de Brachypodium sur un spectromètre MS/MS de type `QTOF`. Nous pouvons noter ici qu'il existe une inégalité de la répartition de l'information (des pics) au sein d'un spectre.

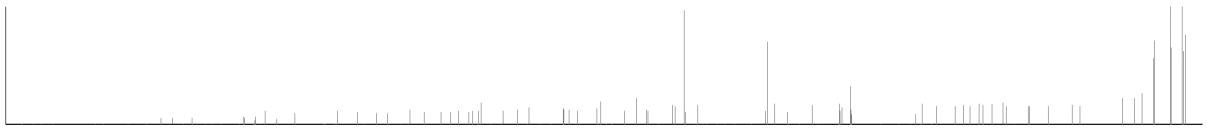

Figure 2.12 – Le spectre MS/MS représente le peptide QQQQQEGEEEGFIIR issu d'une protéine de Brachypodium analysé par un spectromètre de type `QTOF`. Ce spectre illustre la possibilité dans un spectre que la plupart des pics aient une faible intensité.

# CHAPITRE 3

# L'identification de protéines en MS/MS - État de l'art et problématique

## 3.1 Introduction

Ce chapitre décrit le problème de l'identification des protéines à partir d'un ensemble de spectres MS/MS et présente les différentes approches existantes pour interpréter les données issues d'une analyse par spectrométrie de masse en tandem (MS/MS).

En MS/MS, un spectre représente la liste des masses des fragments constituant un peptide d'une protéine analysée. L'identification des protéines se fait en deux étapes. Tout d'abord, il convient de retrouver et d'*associer à chaque spectre* produit par l'analyse d'un échantillon *le peptide correspondant*. Pour cela, nous disposons à la fois de la masse du peptide et d'informations complémentaires liées à la composition en acides aminés du peptide. En combinant les peptides identifiés à partir des spectres expérimentaux, il est ensuite possible de retrouver les protéines qui ont été analysées. Nous parlerons ainsi pour cette seconde étape de *remontée à la protéine*.

Nous allons présenter les deux grandes familles de méthodes permettant d'associer un peptide à un spectre, à savoir : (a) l'interprétation *de novo* et (b) l'identification par comparaison avec des protéines connues. Ensuite, nous détaillerons les difficultés rencontrées lors de l'interprétation des données issues d'organismes non séquencés ou de la prise en compte de certaines, voire de nombreuses, modifications post-traductionnelles.

## 3.2 L'interprétation *de novo* d'un spectre MS/MS

Les méthodes basées sur l'approche dite de séquençage *de novo* cherchent à inférer un peptide à partir des informations contenues dans un spectre MS/MS, et cela sans utiliser l'information contenue dans les banques de protéines connues. Nous allons voir dans un premier temps comment reconstruire manuellement une séquence d'acides aminés à partir d'un spectre, puis, dans un second temps, nous expliquerons les différentes méthodes *de novo* automatisées.

### 3.2.1 L'interprétation manuelle d'un spectre MS/MS

Un spectre peut être interprété en cherchant à reconstruire petit à petit sa séquence d'acides aminés. Cette reconstruction va se faire en calculant la différence de masse entre les pics du spectre. Si une différence de masse entre deux pics correspond à la masse d'un acide aminé, il est possible d'étiqueter cet intervalle avec l'acide aminé. Pour faciliter ce processus, des règles d'interprétation, qui peuvent être dépendantes du spectromètre de masse, sont généralement utilisées. Ces règles vont à la fois permettre de retrouver un pic important qui va servir de point de départ pour l'interprétation complète du spectre, mais vont aussi aider à étiqueter les intervalles mesurés entre les différents pics.

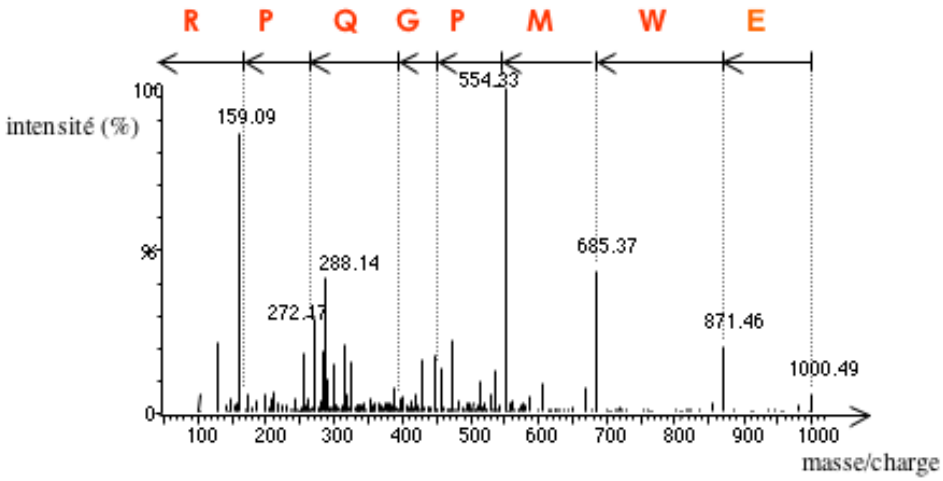


Figure 3.1 – Interprétation *de novo* d'un spectre MS/MS. En interprétant les intervalles entre les pics, il est possible de reconstituer la séquence peptidique. Ici, la séquence peptidique reconstituée se lit de droite à gauche : EWMPGQPR. *(Source : interstices.info)*

Si le spectre étudié est de bonne qualité, il est possible de reconstruire de cette manière la séquence complète du peptide, ainsi que l'illustre la Figure 3.1. Cependant, l'absence de certains pics ou la forte présence de bruit peuvent empêcher de trouver un intervalle correspondant à la masse d'un acide aminé. La limite de l'interprétation est ainsi liée à la qualité des spectres, celle-ci étant représentée par le ratio signal / bruit, dans lequel le signal représente l'information utile et le bruit toute l'information non désirée qui va perturber l'interprétation.

### 3.2.2 L'interprétation automatisée d'un spectre MS/MS

Étant donné la volumétrie des données générées par spectrométrie de masse (un spectromètre en mode MS/MS peut produire plusieurs milliers de spectres par jour), l'interprétation manuelle des spectres trouve très rapidement ses limites. Différentes méthodes ont donc été conçues dans le but d'automatiser l'interprétation des spectres.

Deux grandes familles de méthodes sont utilisées : la méthode pseudo-PFF ou la reconstruction itérative qui repose sur la notion de graphe spectral.

#### 3.2.2.1 Les méthodes pseudo-PFF

À partir de la connaissance de la masse du peptide analysé, ce type d'approche consiste à générer toutes les combinaisons possibles d'acides aminés permettant de retrouver cette masse, et donc à créer une banque de tous les peptides possibles. Un spectre théorique est ensuite généré pour chacun de ces peptides. Les spectres théoriques sont ensuite comparés aux expérimentaux de la même manière que dans une approche PFF (aussi appelée comparaison de spectres), explicité Section 3.3. Le principal défaut de ces méthodes est lié à l'explosion combinatoire du nombre de peptides artificiels générés. Différentes méthodes basées sur ce type d'approche existent et se distinguent par leur manière de limiter l'impact de l'explosion combinatoire.

– L'usage des ions immonium permet de contraindre la composition des peptides. Un **ion immonium** est un fragment présent occasionnellement dans le spectre, et qui signale la présence d'un acide aminé précis sans pour autant donner d'information sur sa position au sein du peptide. Il est donc possible de ne générer que des séquences d'acides aminés contenant les acides aminés désignés par la présence d'ions immonium, ce qui réduit ainsi fortement le nombre de séquences à générer. Spengler et al. ont exploité cette notion dans [Spe04].

– Il est aussi possible de ne pas générer tous les peptides candidats, mais d'utiliser un algorithme génétique pour en générer aléatoirement un certain nombre. Un candidat peut être évalué à l'aide d'une fonction de score. Cette liste de candidats est par la suite modifiée en utilisant des opérations de *recombinaison*, *sélection* et *mutation* qui sont propres aux algorithmes génétiques et qui vont permettre d'obtenir l'ensemble des candidats ayant obtenu les plus hauts scores. Il s'agit de la méthode proposée par Heredia-Langner et al. en 2004 dans [HLCJJ04].

– PEAKS, une méthode développée par Ma et al. en 2003 [MZH$^+$03] fonctionne en deux étapes :
  - Tout d'abord une méthode de programmation dynamique est utilisée pour générer les 10000 meilleurs peptides candidats. Durant cette génération de candidats, le score utilisé ne tient compte que d'assez peu d'informations (abondance des ions a, b, c, x, y ainsi que des variantes b/y présentant une perte de H2O ou de NH3).
  - Ensuite, les 10000 candidats générés sont réévalués en utilisant une méthode de score plus performante, qui va cette fois utiliser une tolérance de masse plus stricte et considérer la présence d'autres types d'ions. Notons que cette limitation n'était pas possible lors de l'étape de programmation dynamique pour des raisons de complexité temporelle.

  Cet outil a pour particularité de donner un score de confiance pour chaque acide aminé en plus de celui donné à la séquence complète.

#### 3.2.2.2 La reconstruction itérative d'une séquence de peptide

Plutôt que de tester un grand nombre de peptides candidats sur chaque spectre expérimental, certaines méthodes tentent de reconstruire le peptide de manière itérative, tout comme dans la méthode manuelle. Ces méthodes s'appuient souvent sur une représentation du spectre sous la forme d'un graphe spectral, qui a été introduit par Bartel en 1990 dans [Bar90].

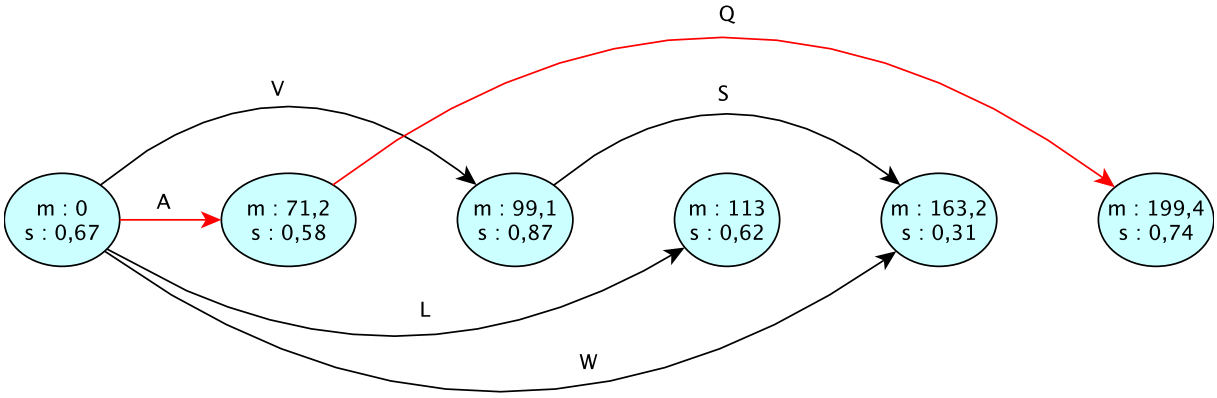


Figure 3.2 – Exemple de graphe spectral. Chaque noeud contient comme information la position (masse m) qu'il représente ainsi qu'un score (s). Plusieurs arcs sont représentés, ils portent tous en étiquette l'acide aminé qu'ils représentent. Chaque chemin de ce graphe représente une séquence d'acides aminés possible. *(Source : présentation de PepNovo par A.M. Frank)*

Dans un **graphe spectral**, tous les ions issus d'un même fragment de peptide sont regroupés en un noeud. Un score est attribué à chaque noeud en fonction de la probabilité qu'une fragmentation ait eu lieu à l'emplacement de ce noeud. Deux noeuds sont reliés par un arc si leur différence de masse est égale à la masse d'un ou plusieurs acides aminés. Chaque arc est étiqueté avec le nom du ou des acides aminés qu'il représente. Il est ensuite possible de rechercher le chemin de score maximal dans ce graphe. Le **score d'un chemin** correspond à la somme des scores de tous les noeuds qu'il emprunte. La lecture des étiquettes le long des arcs empruntés par ce chemin va donner la séquence qui interprète le mieux le spectre. La Figure 3.2 représente un exemple de graphe spectral. Dans ce graphe, le chemin de couleur rouge est le chemin de score maximal. De nombreuses méthodes s'appuient sur cette approche, par exemple SHERENGA [DAC+99], SeqMS [FdCGB+99, FdCGS+00] ou encore LUTEFISK97 [TJ98, TJ01, JT02]. SeqMS propose une méthode pour restreindre les zones du graphe à considérer, permettant ainsi un gain important en terme de temps d'exécution. SHERENGA est une méthode évaluant chaque noeud grâce à un ratio de vraisemblance entre deux hypothèses : (i) les pics sont issus d'une fragmentation, (ii) les pics sont du bruit. En 2005, la méthode SHERENGA a été reprise dans PepNovo avec un calcul de score totalement revu [FP05, FTP05, FSN+07, Fra09b, Fra09a]. Les deux hypothèses utilisées pour calculer le ratio de vraisemblance y ont été modifiées pour prendre en compte de très nombreux aspects, qu'ils soient liés à la séquence protéique, aux conditions expérimentales ou encore à l'appareil utilisé. Pour paramétrer chacun des éléments de ce score, PepNovo utilise un apprentissage.

### 3.2.3 Comparaison des méthodes *de novo*

L'intérêt des graphes spectraux ("spectral graph") par rapport aux méthodes de type Pseudo-PFF réside dans la taille de l'espace de recherche. Les graphes spectraux ont réduit cet espace de

recherche de l'ensemble des peptides possibles à un sous-ensemble correspondant à l'ensemble des chemins de ces graphes.

Ces graphes peuvent cependant présenter des inconvénients, selon la manière dont ils sont utilisés. Généralement, on recherche dans ces graphes un ou plusieurs chemins allant du premier noeud jusqu'au dernier, de manière à obtenir un peptide faisant la même masse que le précurseur. Or, cette hypothèse est parfois trop forte. En effet, dans le cas où certaines fragmentations sont manquantes dans le spectre, il n'est pas possible de trouver un chemin allant du premier au dernier noeud. Dans un tel cas, le graphe contient plusieurs chemins distincts, c'est-à-dire ne partageant aucun sommet, qui représentent chacun une séquence partielle. La majorité des méthodes utilisant des graphes spectraux ne peuvent alors pas traiter de tels cas. Nous pouvons tout de même démarquer PepNovo [FP05] de ces outils, car ce logiciel est capable de produire des tags et donc de ne pas chercher systématiquement à inférer une séquence complète.

De manière plus générale, pour comparer les interprétations de différentes méthodes *de novo*, trois critères sont utilisés :

1. le nombre de séquences retrouvées qui sont identiques à la séquence réelle (employé dans [PFM+06]),
2. le nombre d'acides aminés correctement positionnés qui appartiennent à la séquence réelle (employé dans [BKLL08]),
3. le nombre de tags, d'une taille définie à l'avance, que l'on retrouve dans la séquence réelle (employé dans [PMC07]).

Les critères 1. et 2. permettent d'évaluer les méthodes qui ont pour objectif d'interpréter un spectre complet, c'est-à-dire une très grande majorité. Les différentes comparaisons [BKLL08, PFM+06] tendent à faire ressortir la supériorité de PEAKS [MZH+03] et PepNovo [FP05] sur leurs concurrents.

Le critère 3. ne fonctionne que sur des méthodes capables de fournir des tags en résultat d'interprétation, et non pas une séquence complète. Ce critère est donc bien plus restrictif et ne peut s'appliquer qu'à une minorité de méthodes. Cependant, l'intérêt est fort car un tag, si il est suffisamment long, est suffisant pour confirmer une identification. L'étude de Pitzer [PMC07] montre bien qu'une méthode comme PepNovo est très performante pour fournir des tags à partir de spectres, et que plus la longueur des tags est faible, plus la précision de la méthode est importante, mais l'identification des protéines en devient cependant plus difficile.

## 3.3 L'identification par comparaison avec des protéines connues

### 3.3.1 Le principe général de la comparaison de spectres

Les méthodes reposant sur les banques de protéines consistent à comparer les spectres expérimentaux à des spectres théoriques déduits des protéines contenues dans la banque. En effet, lorsque l'on sait que les protéines recherchées sont référencées dans une banque, il n'est pas nécessaire de rechercher l'interprétation d'un spectre parmi tous les peptides existants, mais uniquement parmi les peptides appartenant aux protéines de la banque de données. Cela réduit ainsi de manière très importante l'espace de recherche à explorer.

Ces méthodes sont appelées méthodes de comparaison de spectres ou PFF (pour Peptide Fragment Fingerprint, par analogie à la méthode PMF décrite Section 2.3.4.2, page 20). Le

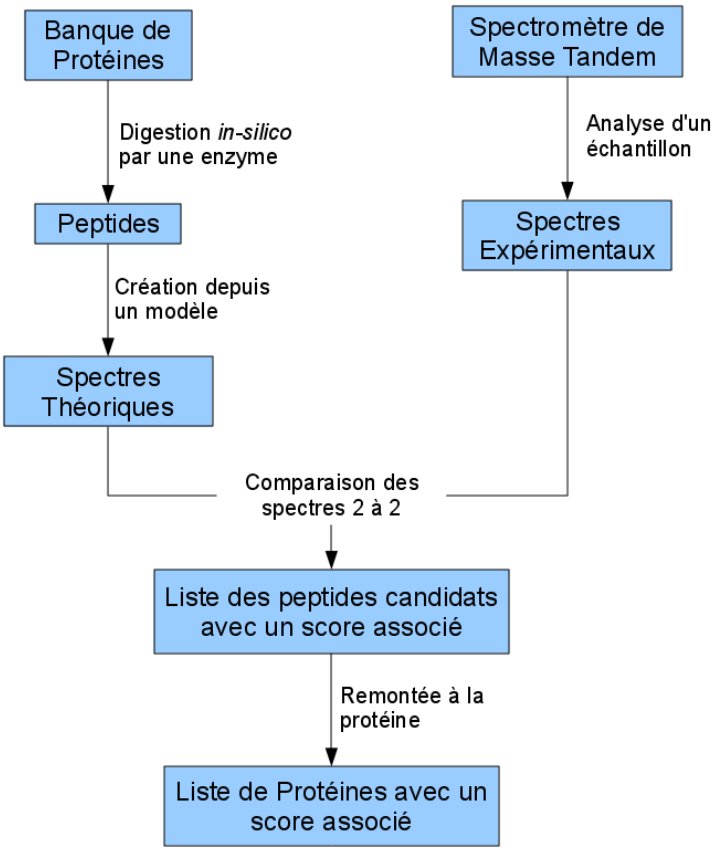


Figure 3.3 – Principe général de la comparaison de spectres (ou PFF).

fonctionnement global de la comparaison de spectres est présenté brièvement dans la Figure 3.3. Le processus de PFF est constitué de différentes étapes commençant par la digestion in-silico des protéines de la banque pour former des peptides, puis ces peptides sont transformés en spectres théoriques dans le but d'être comparés aux spectres expérimentaux produits par le spectromètre de masse. Ces comparaisons vont établir des scores de similarité entre les spectres, ce qui permet d'associer à chaque spectre expérimental, un peptide de la banque (si le score de comparaison avec le spectre théorique de ce peptide est le meilleur). Une fois un ensemble de peptides identifiés, il est possible de rechercher la ou les protéines contenant ces peptides. Nous détaillerons dans les sections à suivre : (i) le choix des peptides candidats à comparer, (ii) la construction de spectres théoriques et (iii) la comparaison de deux spectres avec l'élaboration d'un score qui évalue la similarité entre les spectres.

### 3.3.2 Filtres sur la sélection des peptides théoriques

Les méthodes PFF produisent des peptides en digérant in-silico les protéines d'une banque. Cette digestion va simuler le fonctionnement de l'enzyme utilisée lors de l'expérimentation. On peut noter cependant que certaines méthodes sont capables de rechercher des peptides correspondant à un précurseur donné sans digestion in-silico de la banque. Cependant, cela augmente

considérablement le temps de recherche [LC03].

Pour réduire autant que possible le temps d'identification des peptides et limiter les risques d'associations aléatoires, l'espace de recherche peut être réduit avec l'utilisation de différents filtres. Nous pouvons distinguer deux niveaux de filtrages. Le premier vise à limiter la sélection des protéines de la banque à l'organisme étudié, ou même à choisir une fraction d'un organisme (un chromosome par exemple). Le second niveau de filtrage visera à limiter le nombre de peptides issus de la digestion qui seront considérés pour la comparaison. Il existe différentes manières de filtrer les peptides, la plus commune étant d'utiliser la masse du précurseur. Les méthodes de PFF ne vont donc comparer que des spectres dont les précurseurs ont sensiblement la même masse, la différence de masse tolérée dépendant principalement de la précision de l'appareil utilisé. Ce filtrage sur la masse permet de réduire énormément le nombre de peptides à comparer, mais présente des inconvénients importants. Dans certains cas, la masse du précurseur peut-être erronée, dans d'autres, la différence de masse tolérée peut être insuffisante, et enfin, la présence de modifications peut changer la masse du précurseur. Dans tous les cas, des identifications potentielles peuvent en conséquence être manquées [HHMM10]. Néanmoins, ce type de filtrage reste employé dans pratiquement toutes les méthodes de PFF. Enfin, il est possible d'utiliser des tags obtenus via une interprétation *de novo* pour filtrer les séquences candidates [FTP05, HGFA03, TSYI03]. Un peptide non filtré sera nommé **peptide candidat**.

### 3.3.3 La construction de spectres théoriques

Pour pouvoir comparer les spectres expérimentaux aux peptides candidats, les méthodes PFF doivent construire des spectres théoriques. Ces spectres sont des simulations de la fragmentation des peptides candidats. Généralement, les spectres théoriques sont générés de manière simple. Chaque fragment du peptide va se traduire par la création de plusieurs pics dans le spectre théorique, chaque pic correspondant à un ion différent, ou à une perte neutre. Ces spectres pourraient donc être qualifiés de parfaits au sens où ils présentent toute l'information disponible, et ne contiennent aucun bruit. Un spectre théorique parfait est représenté dans la Figure 3.4 (1.) accompagné d'un spectre expérimental (2.) pour comparaison. Nous pouvons noter une forte différence entre (1.) et (2.), qui vient surtout du fait qu'une partie importante de l'information présente dans un spectre parfait manque dans les spectres expérimentaux, lesquels sont de plus pollués par du bruit.

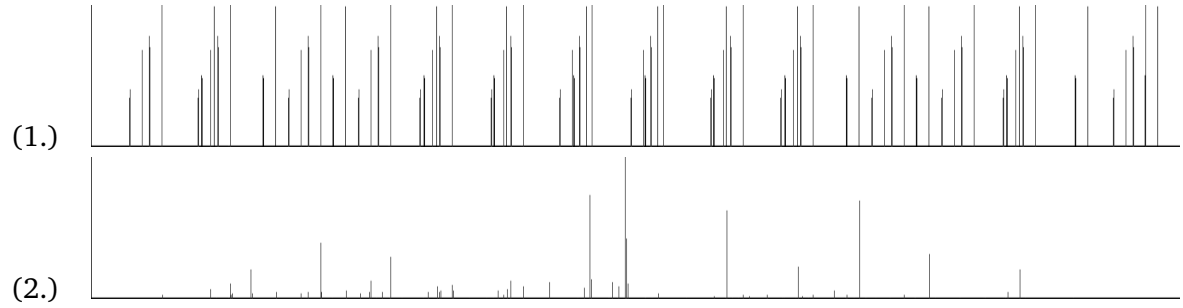


Figure 3.4 – (1.) Exemple de spectre théorique, que nous avons généré, contenant toute l'information utile à l'identification. (2.) Un spectre expérimental issu de l'analyse d'une protéine de Brachypodium à l'aide d'un spectromètre MS/MS de type `QTOF`.

Assez peu de méthodes cherchent à construire des spectres théoriques réalistes, c'est-à-dire, plus proche d'un spectre expérimental que d'un spectre parfait. Pour créer un spectre réaliste, il faut prendre en compte de nombreux éléments lors de la création du spectre, tels que :

- l'influence de l'acide aminé précédant ou suivant une fragmentation,
- la position de la fragmentation dans le peptide (une fragmentation au début ou à la fin ne sera pas marquée de la même manière selon les ions),
- l'influence du type de spectromètre sur la fragmentation (tous les types de spectromètres ne fragmentent pas exactement de la même manière).

En pratique, ces éléments sont très rarement considérés directement au travers du spectre théorique, mais plutôt considérés au travers de la fonction de score, lors de l'évaluation de la similarité entre les deux spectres.

### 3.3.4 Évaluation de la similarité entre deux spectres

Les méthodes de type "comparaison de spectres" ont besoin d'évaluer la similarité entre deux spectres, généralement un spectre théorique et un spectre expérimental. Cette similarité repose sur un modèle de score, qui va avoir pour objectif d'évaluer correctement cette similarité avec si possible un fort pouvoir de discrimination. Il existe plusieurs familles de score en PFF :

- **Nombre de pics en commun :** cette méthode, appelée généralement SPC (pour Shared Peaks Count) utilise un principe semblable a celui employé par la méthode du "nombre de masses partagées" employée en MS. Deux spectres seront considérés comme similaires si ils ont tous deux de nombreux pics situés à la même position. La méthode consiste donc à utiliser comme score, le nombre de pics en commun, c'est-à-dire localisés à la même position ($m/z$) dans les deux spectres. Ce score peut bien évidemment être amélioré en prenant en compte des éléments autres que la présence ou l'absence d'un pic à une position donnée, comme par exemple l'intensité des pics. Ce type de score est le plus utilisé en PFF [GM01], par exemple dans Sequest [EMY94] et Spectrum Mill (application propriétaire vendue par Agilent Technologies).

- **Corrélation croisée :** les deux spectres sont étudiés comme étant des signaux. La corrélation croisée est alors la mesure utilisée pour évaluer la similarité de deux signaux. Cette solution est employée dans Sequest [EMY94] pour recalculer le score des meilleurs candidats trouvés à l'aide du nombre de pics en commun (la corrélation croisée étant plus précise, mais plus longue à appliquer), mais aussi dans Libquest [YIMG+98].

- **Produit scalaire :** les deux spectres sont représentés sous la forme de vecteurs à $N$ dimensions, où $N$ est le nombre de pics communs aux deux spectres. La longueur et la direction d'un vecteur dépendent à la fois des $m/z$ et de l'intensité des pics. Un produit scalaire entre deux vecteurs permet d'obtenir un angle compris entre 0 et 90 degrés. Si l'angle est de 0 degré, les deux spectres sont parfaitement identiques, tandis que si l'angle est de 90 degrés, ils seront considérés comme totalement différents. Le calcul de ce produit scalaire est illustré en Figure 3.5. Cette méthode est par exemple utilisée dans GutenTag [TSYI03], X!Tandem [CB04] ou encore Sonar [FFB02].

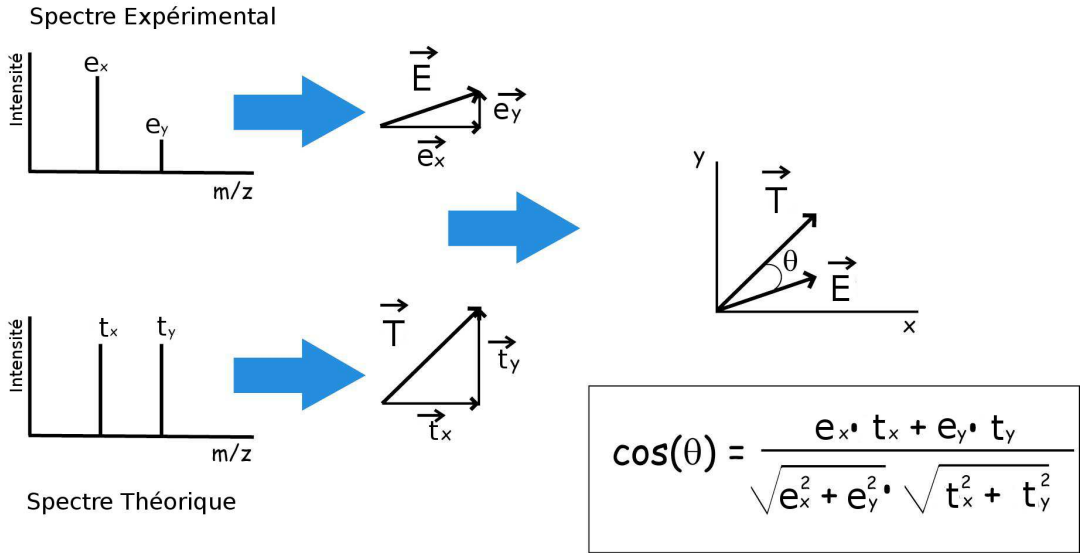


Figure 3.5 – Illustration de la méthode du produit scalaire pour la comparaison de spectres. Cette figure présente la construction du vecteur représentant chacun des spectres, ainsi que la manière dont se calcule le produit scalaire. *(Source : P. Hernandez)*

Toutes ces méthodes sont généralement agrémentées d'un aspect statistique ou probabiliste important qui va garantir la significativité des résultats. C'est par exemple ce qui est effectué dans PepFrag ou Mascot [PPCC99].

### 3.3.5 Bibliothèques Spectrales

Plutôt que de créer des spectres théoriques à partir des séquences protéiques d'une banque, certaines approches proposent de stocker des spectres déjà identifiés dans une banque, puis de les ré-utiliser. L'intérêt de la méthode est multiple : tout d'abord cela permet de réutiliser la connaissance portée par les spectres précédemment identifiés, plutôt que de recréer des spectres théoriques incomplets à chaque comparaison. Ensuite, cela permet un gain de temps et une amélioration des résultats, dans la mesure où tous les spectres contenus dans une banque de spectres sont observables. Nous appelons **observable** un spectre qui peut être produit par un spectromètre de masse en tenant compte de tous les paramètres physiques liés à la protéine analysée ainsi qu'aux contraintes de l'appareil. En effet, dans le cas où les spectres sont générés à partir d'une banque de peptides, nombre d'entre eux ne sont potentiellement pas observables à cause de problème de taille (tolérance du spectromètre de masse, voir Section 2.3.4.1, page 18), d'hydrophobicité (problème lors de l'étape de séparation des données, voir Section 2.3.2, page 14), ou encore de mauvaise découpe enzymatique (voir Section 2.3.3, page 17). Considérer des spectres non observables lors de la comparaison, augmente inutilement les risques de bonne correspondance due uniquement au hasard (aussi appelés *random matches*).

L'approche pionnière est Libquest proposée par Yates et al. en 1998 [YIMG+98]. Cependant, le développement des banques de spectres ayant été plutôt lent, il a fallu attendre 2006 pour voir apparaître d'autres méthodes. X!Hunter, un outil accompagnant X!Tandem a été créé par

Craig et al.[CCFB06]. Frewen et al. ont quant à eux proposé BiblioSpec en 2006 [FMW+06]. SpectraST, développé par Lam et al. [LDE+07], a suivi peu après. Il fut d'ailleurs intégré au Trans Proteomic Pipeline (ou TPP) [KEZ+05], une suite de logiciels spécialisée en protéomique intégrant de nombreux outils.

L'inconvénient majeur de cette méthode est que les banques de spectres sont plus pauvres en contenu que les banques de protéines. Ahrné et al. proposent dans [AMB+09] d'améliorer l'identification en combinant une approche classique de comparaison de spectres (Phenyx) avec une approche de bibliothèque spectrale (SpectraST). L'avantage de leur approche est de ne pas être limitée par le contenu de la banque de spectres.

## 3.4 Comparaison des approches *de novo* et de PFF

Les méthodes permettant l'identification en MS/MS sont nombreuses, mais présentent toutes des défauts plus ou moins importants selon l'usage que l'on désire en faire.

Ainsi, les méthodes *de novo* doivent interpréter un spectre de bonne qualité pour pouvoir correctement déduire une séquence à partir d'un spectre. Des spectres partiels ou avec des défauts (bruit, par exemple) ne donnent que rarement des résultats satisfaisants. Même si les méthodes *de novo* s'améliorent fortement (PepNovo, par exemple, fournit de bons résultats), elles restent dans beaucoup de cas trop peu fiables. Les experts doivent très souvent vérifier manuellement les résultats obtenus via les méthodes *de novo*, voire interpréter manuellement des spectres. Cela a été mis en avant par Pevtsov et al. dans [PFM+06] et confirmé dans [BKLL08, PMC07], au travers d'une comparaison des méthodes *de novo* les plus récentes et les plus employées sur différents jeux de données, et en évualuant différents critères. Il en ressort que les résultats de ces méthodes varient fortement selon les appareils utilisés pour produire les jeux de données, ou encore selon la qualité des spectres. En outre, les tests montrent qu'aucune méthode ne dépasse le seuil de 50% de séquences peptidiques exactes identifiées sur des données de `QTOF`, voire le seuil de 20% de séquences peptidiques exactes pour des données issues d'un spectromètre à piège à ions.

D'un autre côté, la comparaison de spectres présente comme inconvénient sa forte dépendance aux banques de données. Certes, cela est intéressant lorsque l'on désire retrouver un élément de cette banque, mais cela rend la méthode dépendante de la taille de cette banque, et peut induire des temps d'exécution importants. Cela est d'autant plus vrai si tous les filtres ne peuvent être appliqués pour choisir les peptides candidats (par exemple lorsqu'on recherche des modifications, ou que les spectres proviennent d'un organisme inconnu).

## 3.5 La problématique des modifications sans a priori

### 3.5.1 La protéomique des organismes non séquencés

Beaucoup d'organismes d'intérêt ne sont toujours pas séquencés à ce jour, malgré l'important développement des techniques de séquençage. Si nous pouvons dénombrer plus de 1350

génomes complets séquencés ou en cours de séquençage aujourd'hui, moins de 10% sont des eucaryotes (environ 130 organismes), et moins de 10% de ces eucaryotes sont des plantes.

Finalement, fin 2010 seulement 12 plantes différentes sont séquencées, dont 6 en 2009-2010 (cf. Table 3.1). Ce nombre est très faible par rapport aux 260.000 espèces de plantes référencées dans le monde [JRP06].

| **Nom** | **Nom commun** | **Année de séquençage** |
|---|---|---|
| Arabidopsis Thaliana | Arabette des dames | 2000 |
| Oryza Sativa | Riz | 2002 |
| Physcomitrella Patens | Bryophyte | 2002 |
| Populus Trichocarpa | Peuplier | 2006 |
| Vitis Vinifera | Vigne | 2007 |
| Carica Papaya | Papaye | 2008 |
| Cucumis Sativus | Concombre | 2009 |
| Zea Mays | Maïs | 2009 |
| Sorghum Bicolor | Sorgho | 2009 |
| Glycine Max | Soja | 2010 |
| Brachypodium Distachyon | | 2010 |
| Malus Domestica | Pommier | 2010 |

Table 3.1 – Liste des plantes séquencées à ce jour.

Une façon de procéder à l'identification de protéines d'organismes non séquencés consiste à utiliser des protéines homologues. Ces protéines homologues proviennent d'organismes proches (en termes d'évolution des espèces) de celui qui est étudié, et doivent être parfaitement connues.

Deux protéines sont dites **homologues** si elles sont le résultat d'un processus d'évolution divergeant à partir d'un ancêtre commun [Fit00]. Il existe plusieurs types de protéines homologues :

- les **protéines orthologues**, qui sont issues d'une spéciation (deux protéines orthologues appartiennent donc à deux espèces différentes), et
- les **protéines paralogues**, qui sont issues d'une duplication au sein d'une même espèce.

En protéomique des organismes non séquencés, il est donc fréquent de rechercher des protéines orthologues parmi les protéines d'une espèce connue. Cela permet d'en extraire de la connaissance, telle que sa famille protéique ou encore sa fonction, qui va être réutilisable sur la protéine non séquencée analysée.

### 3.5.2 Différents types de modifications des protéines

Les modifications des protéines sont des changements fondamentaux au niveau des acides aminés. Nous pouvons les considérer de différents types, selon que l'on ait une connaissance a priori ou non de ces modifications. Lorsque l'on en a une connaissance a priori, c'est-à-dire une supposition de leur présence, nous les classons généralement en deux types :

- Les **modifications fixes** qui transforment simplement toutes les occurrences d'un acide aminé en une entité différente (un autre acide aminé, ou un acide aminé avec un composé chimique qui lui est greffé). En pratique, une telle modification peut simplement être

vue comme un changement de masse de l'acide aminé modifié. Ces modifications sont généralement volontairement créées lors des manipulations. Par exemple, la carbamidométhylation des cystéines est une conséquence de la suppression des ponts disulfures, une modification de la structure des peptides qui gêne l'analyse par spectrométrie de masse en empêchant une fragmentation correcte des peptides. Cette modification induit un changement de masse de $+57$ Da sur les cystéines de la protéine.

– Les **modifications variables** touchent également un acide aminé, mais pas sur toutes ses occurrences, uniquement sur certaines. Pour chacune des occurrences de cet acide aminé, il faudra donc considérer la possibilité qu'il ait changé ou non. La majorité des modifications post-traductionnelles est variable. Par exemple la phosphorylation est une modification post-traductionnelle très courante, qui intervient principalement sur les sérines et les thréonines avec l'ajout d'un groupe phosphate. Cette modification est très connue pour être un élément de régulation : elle va permettre d'activer ou de désactiver de nombreux enzymes et récepteurs.

En revanche, si nous n'avons aucun a priori sur les modifications, c'est-à-dire si on ne sait pas quelles modifications sont potentiellement présentes ou quels changements les modifications induisent (dans le cas d'une modification non encore référencée dans les banques), nous parlerons de **modifications sans a priori**. Nous noterons que dans le cadre de la comparaison de protéines d'organismes non séquencés avec des banques de protéines homologues, il est nécessaire d'envisager des modifications sans a priori. Nous pouvons également considérer que la prise en compte de très nombreuses modifications post-traductionnelles conduit à la même problématique qu'une approche sans a priori au vu de l'explosion combinatoire qu'elle génère.

### 3.5.3 Conséquences d'une modification dans un spectre

Nous avons défini dans le paragraphe précédent les types de modifications pouvant exister, ainsi que leur impact sur les séquences d'acides aminés. Mais nous pouvons aussi observer l'importance des changements qu'elles engendrent dans un spectre. En effet, modifier un seul acide aminé parmi tous ceux composant un peptide change considérablement l'apparence du spectre qui le représente : c'est ce qui rend l'identification de spectres en présence de modifications difficile. La Figure 3.6 illustre ce phénomène sur les spectres qui représentent respectivement le peptide PRTEIN et le peptide PRTEYN. Les deux spectres présentent de fortes différences : en effet, si ces deux peptides ont cinq acides aminés sur six en commun, les spectres n'ont quant à eux que cinq pics sur dix qui soient localisés aux mêmes positions. Cet important changement au sein du spectre rend les méthodes d'identification avec modifications plus compliquées.

### 3.5.4 Comparaison des différentes approches d'identification en présence de modifications

#### 3.5.4.1 Problème combinatoire avec les méthodes PFF

Les méthodes de PFF comparent traditionnellement les spectres expérimentaux à des spectres théoriques générés depuis les données issues d'une banque. S'il y a des modifications dans le peptide, cette comparaison est rendue plus difficile, car un certain nombre de pics ne seront plus

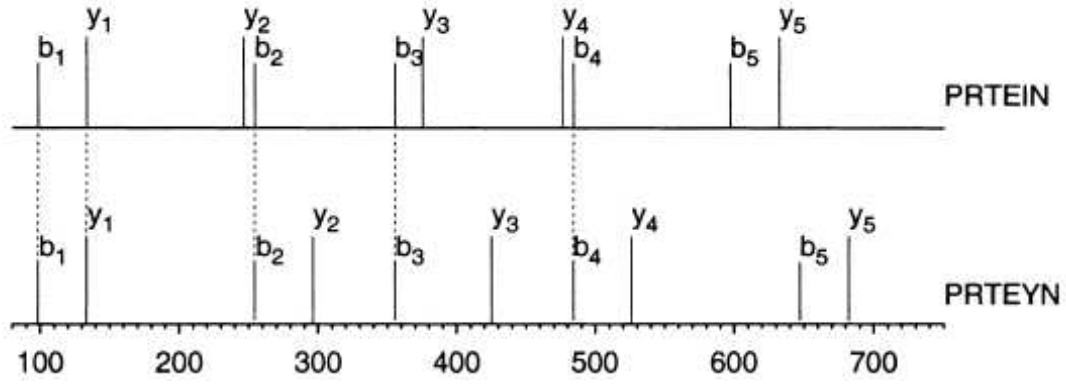


Figure 3.6 – Cette figure illustre, à l'aide de spectres simplifiés, à quel point une modification d'acide aminé peut changer l'apparence d'un spectre. A la droite de chaque spectre se trouve la séquence d'acides aminés qu'il représente. En pointillés, les 5 pics présents aux mêmes positions. *(Source : Pevzner [PMDT01])*

alignés. De plus, ces méthodes filtrent généralement les banques sur la masse du précurseur ; or, toute modification fait varier cette masse. Il n'est donc plus possible de filtrer les données de cette manière. Une des possibilités pour gérer les modifications en comparaison de spectres consiste à utiliser une liste de modifications, ce qui nécessite de connaître à l'avance toutes les modifications possibles. Les méthodes de comparaison vont donc générer toutes les variantes modifiées pour chaque peptide en utilisant une liste des modifications possibles. Certaines méthodes, comme Phenyx [CMG+03], InsPect [TSF+05] ou Mascot [PPCC99] permettent de traiter ces modifications connues a priori, qu'elles soient fixes ou variables.

**Modification fixe.** Une modification fixe est facilement prise en compte par les méthodes de PFF. Les peptides sont simplement modifiés avant de créer le spectre théorique. Il n'y a donc pas d'augmentation du nombre de candidats.

**Modification variable.** Une modification variable n'est pas présente systématiquement dans les peptides. Pour les prendre en compte, les méthodes de type PFF doivent générer toutes les combinaisons possibles de peptides modifiés. Cela va augmenter de manière importante le nombre de candidats à comparer (voir par exemple la Figure 3.7), et par conséquent le temps d'exécution. Il n'est donc pas possible de gérer tous les types de modifications variables de cette manière, mais uniquement une liste réduite.

Comme il n'est pas possible de rechercher une grande variété de modifications variables simultanément via des méthodes de type PFF, certaines stratégies alternatives ont été développées. Par exemple, Phenyx [CMG+03], Mascot [PPCC99], X !Tandem [CB04] ou encore VEMS [MLWB03, MBS+04, MTH+05, Mat07b] effectuent :

– une première identification en ne recherchant qu'un faible nombre de modifications dans le but d'identifier des protéines, puis

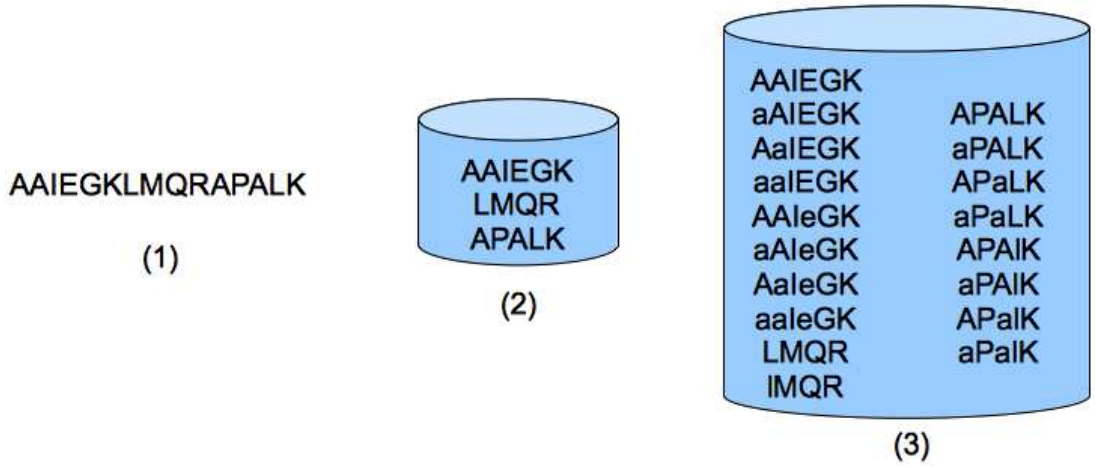


Figure 3.7 – Nous illustrons ici le problème de combinatoire qui se pose lorsque la banque de protéines doit être modifiée pour prendre en compte de nombreuses modifications variables. La banque (2) contient les peptides issues de la protéine (1) lorsqu'aucune modification n'est attendue, tandis que la banque (3) contient les peptides lorsque l'on a un a priori sur 3 modifications variables (substitutions respectives de **A** en **a**, **E** en **e** et **L** en **l**).

– un second passage sur les protéines trouvées précédemment en recherchant un nombre plus important de modifications.

Cette approche suppose qu'au moins un peptide non modifié ou peu modifié -en fonction du paramétrage de la méthode- a permis d'identifier la protéine analysée.

Cette technique fonctionne très bien pour rechercher des modifications dont on soupçonne l'existence, mais elle est difficilement applicable sur une recherche sans a priori. Dans un tel cas, en supposant que l'on connaisse l'existence de toutes les modifications possibles (ce qui n'est pas le cas), il faudrait générer un nombre beaucoup trop important de candidats. C'est pour cette raison que des méthodes ont été créées spécifiquement pour rechercher des modifications sans a priori.

#### 3.5.4.2 Des approches plus adaptées aux modifications sans a priori

**Tags.** Les approches par tags sont apparues en 1994 grâce à Mann et Wilm [MW94]. L'idée est d'utiliser une méthode de type *de novo* pour interpréter seulement les parties de bonne qualité de chacun des spectres. Mann et Wilm appellent tag une courte séquence d'acides aminés associée à deux masses, la masse des résidus précédant et la masse des résidus succédant cette séquence.

Après avoir créé des tags pour chacun des spectres, une recherche de motifs est effectuée dans la banque. L'objectif est de trouver des peptides partageant la séquence du tag et pour lesquels au moins une des deux masses de résidu associées est correcte. Ainsi, si une des deux masses n'est pas correcte, il est possible de supposer qu'il y a modification. Si les deux masses de résidus correspondent parfaitement, alors il n'y a pas de modification. Dans le cas où il y a modification, seule la partie non modifiée est alignée avec une méthode de type PFF. Aucune

information n'est utilisée pour valider l'éventuelle modification. Un exemple de tag accompagné d'une proposition d'alignement est présenté en Figure 3.8.


Figure 3.8 – Sélection d'un candidat à l'aide d'un tag. Le tag (représenté en bas) contient une courte séquence d'acides aminés accompagnée de deux masses. La correspondance proposée ici présente une modification de 100 Da dans la dernière partie du peptide. *(Source : Hernandez [Her05])*

Cette approche a été reprise et améliorée par Tabb et al. en 2003 avec GutenTag [TSYI03]. Cette méthode utilise une extraction de tag automatique, soutenue par une méthode de score beaucoup plus sophistiquée que celle de Mann et Wilm. GutenTag a permis d'automatiser le fonctionnement de l'approche par tags.

Une autre manière d'utiliser les tags a été proposée par Searle et al. en 2004 avec OpenSea [SDT+04, SDW+05]. Leur méthode utilise en entrée les séquences *de novo* fournies par l'application de PEAKS. OpenSea extrait des tags depuis ces séquences, puis va rechercher dans une banque tous les peptides les contenant. Ensuite, cette méthode cherche à étendre l'alignement entre la séquence produite par PEAKS, et les peptides candidats. Il s'agit d'un alignement de séquence local basé sur la masse de groupes allant de 1 à 3 acides aminés. Si aucun alignement n'est trouvé pour un groupe, alors la méthode en déduit une modification. Si une modification est détectée, alors le groupe suivant ne peut en contenir.

Cette approche a fortement inspiré Frank et al. [FP05] dans la création d'un filtre sur la base de tags. La principale différence avec GutenTag vient du fait que les candidats ne sont plus alignés localement en utilisant les séquences, mais en utilisant les spectres. Ainsi, les parties des spectres contenant le tag sont alignées, puis l'alignement est étendu des deux côtés en autorisant des modifications. Cette solution est utilisée dans InsPecT [TSF+05]. En revanche, si InsPecT améliore effectivement l'usage des tags dans cette approche, il ne permet plus l'identification de modifications sans restriction, il nécessite en effet une liste de modifications.

Enfin, Popitam [HGFA03] propose d'extraire des tags via une méthode *de novo*, puis en s'appuyant sur une banque de protéines, crée des scénarios d'interprétation. Un scénario correspond à une manière de positionner différents tags potentiellement séparés par des *gaps* (espace de taille connue, mais de contenu inconnu). Chaque gap est ensuite interprété pour en déterminer la cause (modification, manque d'information dans le spectre, etc.). Les scénarios sont ensuite évalués en fonction de leurs gaps, selon qu'ils présentent ou non des modifications. Un seul scénario est conservé pour chaque spectre.

Les approches par tags sont très intéressantes, en particulier pour leur rapidité d'exécution. Elles sont particulièrement adaptées pour détecter un faible nombre de modifications. Cependant, dans le cas où l'on travaille sur un organisme proche, les mutations peuvent être bien plus fréquentes que les modifications post-traductionnelles, et le risque qu'une substitution, insertion ou suppression d'acide aminé soit présente dans le tag n'est pas négligeable. De plus, les approches par tags dépendent fortement de la qualité de la méthode *de novo* employée afin d'obtenir les tags. Or, il a été montré à plusieurs reprises que le *de novo* peut présenter des faiblesses [PFM+06, PMC07].

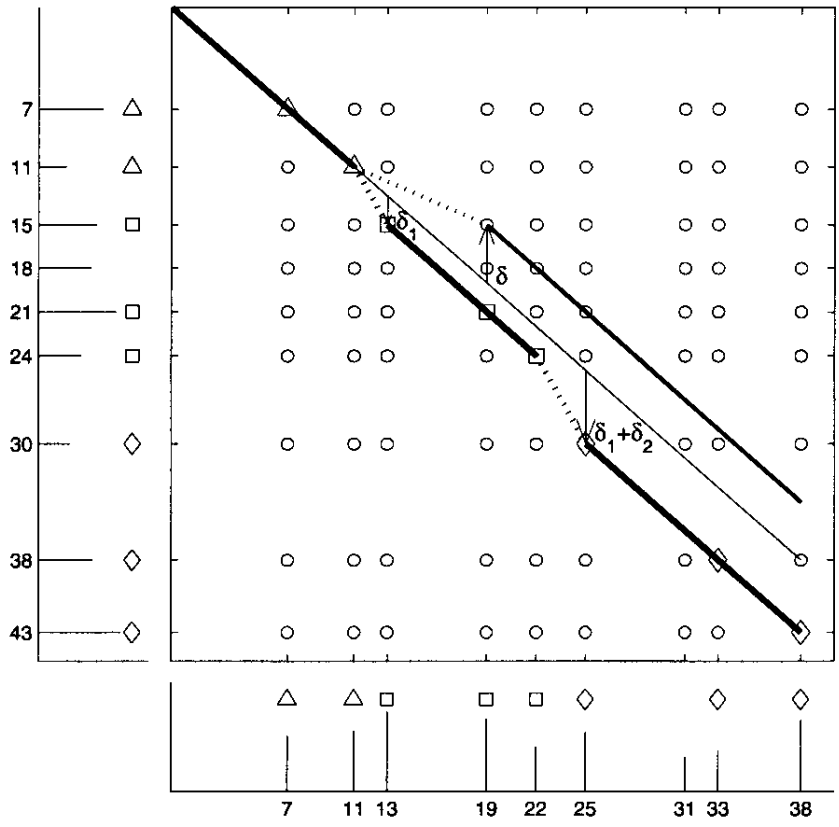


Figure 3.9 – Alignement de $V_t$ avec $V_e$ dans SA. Le nombre de points (symbolisés ici par les symboles △, □, ◇ et ○) situés sur la diagonale principale correspond au score SPC sans aucune modification (ici le score serait donc de 3). En tolérant des modifications, SA autorise des changements de diagonale (correspondant à des modifications notées $\delta$). En tolérant une unique modification $\delta$, il existe un alignement de score 5. Si l'on tolère deux modifications $\delta_1$ et $\delta_2$, il existe un alignement de score 8 (l'alignement passe par les △, puis par les □ après la première modification $\delta_1$ et enfin par les ◇ après la seconde modification $\delta_2$). *(Source : Pevzner et al. [PDT00])*

**SpectralAlignment.** La méthode SpectralAlignment (SA) a été développée par Pevzner et al. [PDT00, PMDT01]. Cette approche permet la recherche de modifications sans a priori. Sa

particularité est de reposer exclusivement sur une approche de type comparaison de spectres et non pas *de novo*, à l'inverse de tous ses concurrents. SA est une méthode qui aligne deux spectres en autorisant des décalages de pics afin d'obtenir le meilleur alignement.

En considérant que chacune des masses pour lesquelles un pic est observable peut être vue comme un entier, un spectre est alors représenté par un vecteur de booléens. Ce vecteur contient, pour chacune des masses possibles, un booléen témoignant de la présence d'un pic ("vrai") ou de son absence ("faux"). Soit $V_t$ le vecteur représentant le spectre théorique et $V_e$ le vecteur représentant le spectre expérimental. SA aligne, deux à deux, les éléments de $V_t$ avec les éléments de $V_e$, en autorisant l'insertion de gaps comme dans le cas de l'alignement de séquences d'acides aminés. Il est juste nécessaire de s'assurer que les deux vecteurs aient la même longueur, et qu'un gap n'est pas aligné avec un autre gap. Dans cette représentation, une modification entre un spectre théorique et un spectre expérimental qui se traduit par un décalage de pics va correspondre à l'insertion d'un gap dans $V_t$ ou $V_e$.

Une fonction de score est utilisée pour évaluer l'alignement, c'est-à-dire la similarité entre les deux spectres. Un score plus élevé signifie une plus grande similarité. Par exemple, la fonction de score Shared Peaks Count (SPC) introduite en Section 3.3.4, page 32 et transposée aux vecteurs $V_t$ et $V_e$ correspond au nombre de paires de booléens à "vrai" situés à la même position.

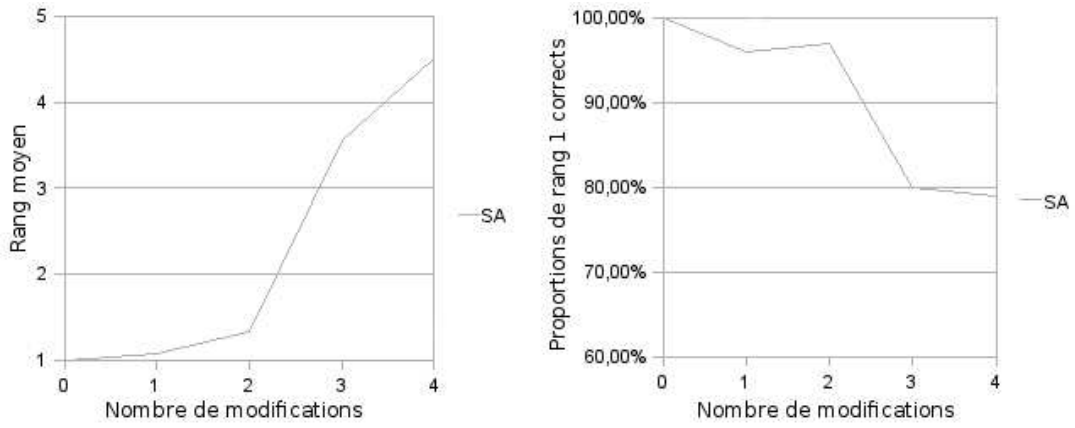


Figure 3.10 – Comportement de SpectralAlignment en fonction du nombre de modifications présentes sur un jeu de 1000 spectres créés *in silico*. Le rang représente ici le classement de l'identification correcte, parmi tous les peptides candidats ordonnés par score décroissant. Il est visible ici qu'avec plus de deux modifications, le taux d'identification décroît considérablement. *Plus de détails sont disponibles dans l'Annexe A, page 131.*

La méthode SpectralAlignment utilise la programmation dynamique pour trouver le meilleur alignement possible entre un spectre théorique et un spectre expérimental (voir l'illustration qui en est donnée à la Figure 3.9). L'approche proposée ici favorise la prise en compte de modifications sans a priori au détriment des temps d'exécution. Dans sa version initiale, le système de score proposé est très simple et est assez peu approprié aux jeux de données réels. La réalisation

de tests présentés à la Figure 3.10 montre que cette méthode est adaptée à la prise en compte d'une ou deux modifications par peptide, mais trouve très rapidement ses limites si le nombre de modifications lui est supérieur. Pour comprendre l'origine de cette limitation, nous allons reprendre notre exemple de spectres correspondant aux peptides PRTEIN et PRTEYN présentés Figure 3.11. Un troisième spectre détaille comment SpectralAlignment modifie le spectre du peptide PRTEIN pour l'aligner avec celui du peptide PRTEYN. La flèche bleue représente le décalage des pics qui permet d'aligner deux pics supplémentaires. SpectralAlignment permet donc d'obtenir l'alignement de 7 pics sur 10, contre uniquement 5 sur 10 sans réajustement de l'alignement. Enfin, le quatrième spectre représente le peptide PWTEYN, qui diffère donc de deux acides aminés comparativement au peptide PRTEIN. Ce quatrième spectre illustre qu'un alignement, après deux modifications, est difficile. En effet, un simple SPC n'alignera que 2 pics sur les 10, mais le meilleur alignement obtenu par SpectralAlignement ne donnera au mieux qu'un score de 5 pics alignés sur 10. Cela nous montre bien à quel point la mise en correspondance des pics devient difficile en présence de plusieurs modifications, et SpectralAlignment trouve rapidement ses limites dans ce cas.

SpectralAlignment a été repris et fortement amélioré dans MS-Alignment par Tsur et al. [TTZ$^+$05], mais ces améliorations restreignent la méthode à l'identification de modifications post-traductionnelles. MS-Alignment permet de détecter un faible nombre de modifications, et s'appuie sur une matrice de fréquence d'observation des modifications pour calculer le score, afin de valider ou non leur présence. Cette méthode n'est donc plus viable pour rechercher des modifications sans a priori. Par la suite, MS-Alignment a été intégré dans InsPecT [TSF$^+$05], où il a été couplé à une approche par tag, déjà évoquée ci-avant.

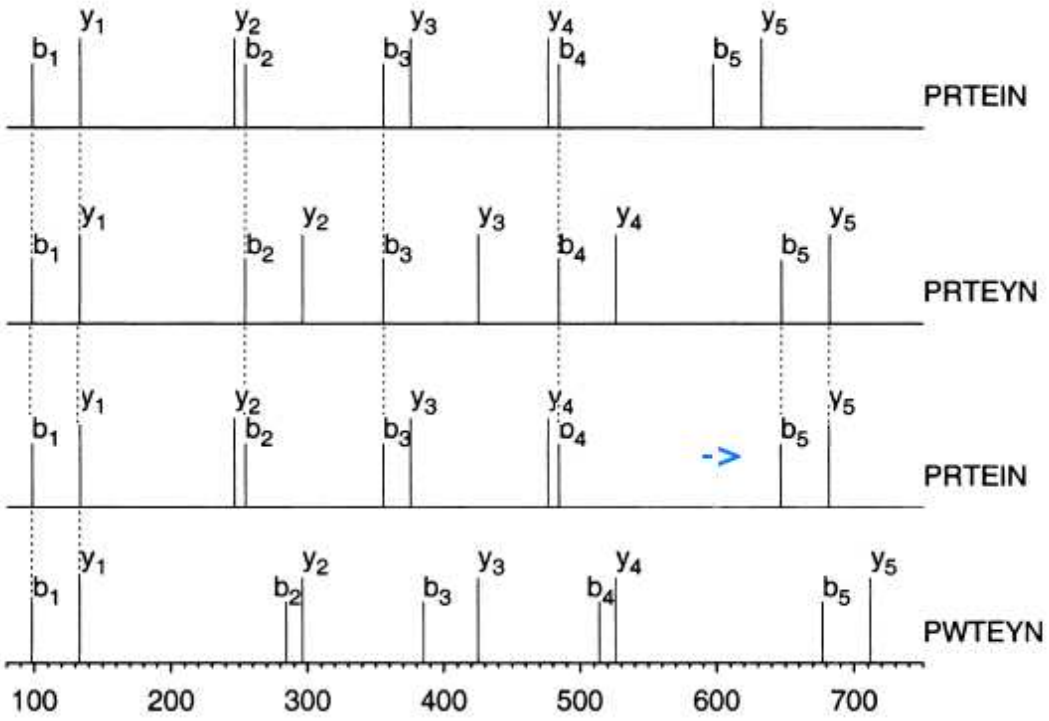


Figure 3.11 – Ces spectres simplifiés illustrent l'impact des modifications sur l'alignement de deux spectres, et l'apport de SpectralAlignment, ainsi que ses limites. Les deux premiers spectres présentent l'alignement du spectre du peptide PRTEIN avec le spectre représentant le peptide PRTEYN. Le troisième spectre montre l'alignement des mêmes spectres que précédemment, mais avec le décalage des pics opéré par SpectralAlignment. La flèche bleue montre ce décalage et permet d'aligner deux pics supplémentaires. Enfin, le dernier spectre représentant le peptide PWTEYN illustre le fait qu'en présence de plusieurs modifications (2 dans ce cas) nous obtenons un score très faible (2 pics alignés sur 10 dans ce cas). *(Source : Pevzner [PMDT01])*

# CHAPITRE 4

# PacketSpectralAlignment, une nouvelle méthode de comparaison de spectres

## 4.1 Introduction

Permettre l'identification de protéines dans le cas d'organismes non séquencés reste un problème difficile. Les méthodes de type *de novo*, bien que majoritairement employées, présentent de forts problèmes en terme de qualité des résultats. Cependant, la seule alternative est la comparaison de spectres, qui, même si elle fournit généralement des résultats de bien meilleure qualité, ne propose pas de méthodes satisfaisantes à l'heure actuelle. Comme nous avons pu le constater dans le chapitre précédent, les méthodes de comparaison de spectres sont limitées :

- soit par un problème de complexité algorithmique
- soit par le nombre de modifications tolérées lors de la recherche, généralement limité à deux.

Nous avons donc décidé de concevoir une nouvelle méthode basée sur la comparaison de spectres, afin de bénéficier de ses avantages en terme de qualité de résultats, tout en prenant soin d'autoriser la présence éventuelle de plusieurs (deux ou plus) modifications. Cette méthode devra en outre permettre une identification de modifications sans a priori, et ce dans un temps d'exécution raisonnable.

Nous introduirons tout d'abord des notations relatives aux spectres ; ensuite nous définirons deux nouvelles notions, la symétrie interne aux spectres et les paquets. Ces notions vont nous permettre de réaliser des alignements de meilleure qualité en présence de modifications. Enfin, nous pourrons détailler leur usage avant de décrire l'algorithme d'alignement utilisé pour exploiter ces notions. Ces travaux, réalisés en collaboration avec Guillaume Fertin, Irena Rusu et Dominique Tessier, ont été publiés dans [CFRT09].

## 4.2 Notations

Soit $P$ un peptide constitué de $N$ acides aminés. Chacun de ces acides aminés sera noté $aa_i$ avec $i \in [1; N]$.

**Masse d'une séquence d'acides aminés.** La masse d'une séquence d'acides aminés $S$ est notée $m(S)$. Ainsi la masse de $P$ est $m(P)$ et $m(aa_1...aa_i)$ représente la masse de la séquence constituée des acides aminés $aa_1$ à $aa_i$.

**Séquences complémentaires dans $P$.** Dans un peptide $P$ de taille $N$, la séquence constituée des acides aminés $aa_1$ à $aa_i$ a pour **complémentaire** la séquence constitué des acides aminés $aa_{i+1}$ à $aa_N$.

**Fragmentation du peptide.** Une **fragmentation** de $P$ dans le spectromètre de masse en mode MS/MS après l'acide aminé $aa_i$ va engendrer de nombreux fragments. Ces fragments sont nommés **N-terminaux** si ils sont constitués des acides aminés $aa_1$ à $aa_i$, **C-terminaux** si ils sont constitués des acides aminés $aa_{i+1}$ à $aa_N$.

**Ions.** Les fragments d'un peptide dont le $m/z$ est évalué par le spectromètre de masse ont été ionisés dans cet appareil et sont donc appelés **ions**. Les **ions N-terminaux** les plus fréquents sont nommés $x$, $y$ et $z$. De manière similaire, les **ions C-terminaux** les plus fréquents sont nommés $a$, $b$ et $c$. La Figure 4.1 met en correspondance les différents fragments ($a$ avec $x$, $b$ avec $y$ et $c$ avec $z$) en fonction des différents sites de fragmentation. On dira que l'ion $b$ est l'ion **complémentaire** de l'ion $y$, et inversement.

Chacun de ces ions peut subir des **pertes neutres**, parmi lesquelles on trouvera souvent une perte en eau ($H_2O$) ou une perte en ammoniac ($NH_3$), formant ainsi un nouvel ion. Un ion $b$ subissant une perte d'$H_2O$ forme un ion $b^\circ$, tandis que si il subit une perte de $NH_3$ il forme un ion $b^*$.

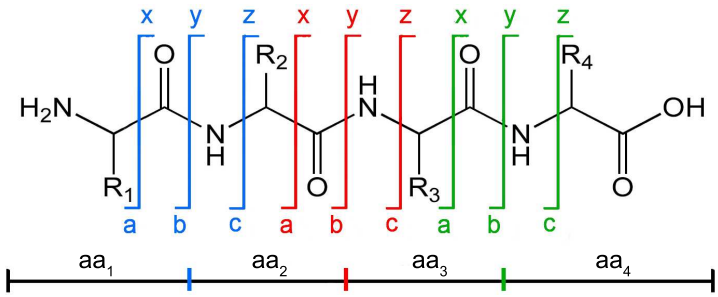


Figure 4.1 – Les différents sites de fragmentation à l'intérieur d'un peptide de $N = 4$ acides aminés ($aa_i$ avec $i \in [1;4]$). $R_i$ ($i \in [1;4]$) représente la chaîne latérale de $aa_i$ (voir Section 2.2.1.2, page 8).

**Masse d'un ion.** Un ion a une masse qui correspond à la somme des masses des acides aminés le constituant, à laquelle s'ajoute un delta, positif ou négatif, dépendant du type d'ion. Ainsi, par exemple, un ion $b^\circ$ aura une masse correspondant à la somme des masses de ses acides aminés minorée par un delta représentant la masse de l'eau. Nous désignerons par $\delta_\alpha$ le delta de masse correspondant à un ion $\alpha$. Ainsi, nous avons par exemple $\delta_{b^\circ} = 18$, le delta correspondant à l'ion $b^\circ$.

**Pic.** Le spectre de masse d'un peptide $P$ est représenté par une suite ordonnée de pics. Chaque pic $p$ est caractérisé par une abscisse, notée $m(p)$, qui correspond à la masse divisée par la charge de l'ion qu'il représente ($m/z$) et une intensité qui est liée à l'abondance de l'ion dans l'analyseur. Un pic correspondant à un ion N-terminal (respectivement C-terminal) sera appelé **pic N-terminal** (respectivement **pic C-terminal**). Pour une fragmentation entre les acides aminés $aa_i$ et $aa_{i+1}$, nous noterons $x_i$, $y_i$ et $z_i$ les pics qui représentent les ions N-terminaux, et $a_i$, $b_i$ et $c_i$ les pics représentant les ions C-terminaux.

Nous supposerons généralement que la charge d'un ion est 1. Ramener cette charge à 1 est un traitement fréquent sur les spectres, qui permet de simplifier le ratio $m/z$. Cela explique qu'il est souvent fait mention de la masse plutôt que du $m/z$ pour qualifier la position d'un pic.

## 4.3 Deux notions importantes : Symétrie et Paquets

L'algorithme original, PacketSpectralAlignment, que nous avons développé dans le cadre de cette thèse exploite deux caractéristiques particulières des spectres de masse en tandem que nous allons maintenant détailler. Il s'agit d'une part de la symétrie interne et d'autre part de la notion de paquets.

### 4.3.1 La Symétrie interne au spectre

Il existe une **notion de symétrie** dans les spectres MS/MS entre les pics N-terminaux et les pics C-terminaux. Comme nous l'avons vu précédemment, la fragmentation d'un peptide $P$ après l'acide aminé $aa_i$ donne lieu à la création de différents ions N-terminaux et C-terminaux. Parmi ces ions N-terminaux, nous pouvons en choisir un que nous nommerons $\alpha$, puis nous définissons $\beta$ comme l'ion C-terminal complémentaire de $\alpha$. Les pics $\alpha_i$ et $\beta_i$, représentant ces deux ions dans le spectre, sont liés par l'Équation 4.1.

$$m(\alpha_i) = m(P) - m(\beta_i) + C \tag{4.1}$$

Dans cette équation, $C$ est une constante que l'on ajoute en raison du fait que les peptides ne sont pas symétriques à leurs extrémités (un phénomène visible dans la Figure 4.1), ainsi qu'à l'ionisation. Ce qu'il est important de noter, c'est que cette notion de symétrie est valide pour toutes les fragmentations issues d'un peptide.

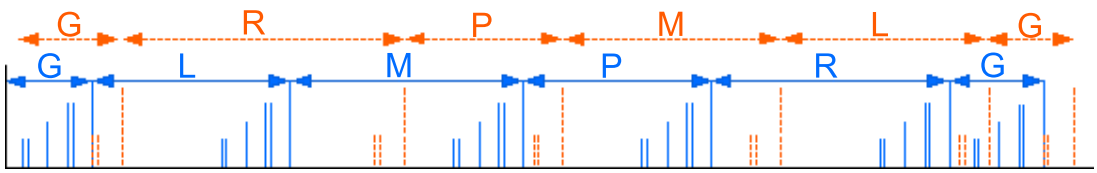


Figure 4.2 – Spectre MS/MS avec en bleu les pics représentant les ions N-terminaux accompagné de la séquence d'acides aminés qu'ils constituent, et en orange les pics représentant les ions C-terminaux accompagnés de la séquence d'acides aminés qu'ils constituent.

Cette symétrie des pics est directement visible dans le spectre. En effet, la série de pics N-terminaux permet de déduire la séquence d'acides aminés du peptide analysé en interprétant les différences de masse de gauche à droite dans le spectre (visible en orange dans la Figure 4.2) ; tandis que la série de pics C-terminaux permet de déduire la séquence d'acides aminés de droite à gauche (visible en bleu dans la Figure 4.2). Dans l'Équation 4.1, $m(\alpha_i)$ correspond aux pics N-terminaux, tandis que $m(\beta_i)$ correspond aux pics C-terminaux (toujours en supposant que $\alpha$ est l'ion complémentaire de $\beta$).

### 4.3.2 Les Paquets

Nous avons vu qu'un peptide se fragmente en de nombreux ions, et qu'il existe une relation de symétrie entre certains pics représentant ces fragmentations. Nous allons introduire la notion de **paquets** pour représenter la cohérence qui existe entre les pics issus d'une même fragmentation. Ainsi, le paquet $p_i$ contiendra tous les pics issus d'une fragmentation entre les acides aminés $aa_i$ et $aa_{i+1}$ du peptide $P$, c'est-à-dire tous les pics représentant les ions constitués de la séquence d'acides aminés $aa_1$ à $aa_i$ et de sa séquence complémentaire.

La Figure 4.3 représente un spectre MS/MS dans lequel les différents paquets sont identifiables à l'aide de couleurs différentes. Il y a dans ce spectre, qui représente le peptide GLMPRG, 7 paquets différents, séparant chacun des acides aminés du peptide. Nous pouvons noter que les paquets marquant les deux extrémités contiennent moins de pics que les autres paquets : seuls les N-terminaux ou les C-terminaux sont présents selon l'extrémité concernée. Un spectre représentant un peptide $P$ de longueur $N$ est en conséquence constitué de $N + 1$ paquets.

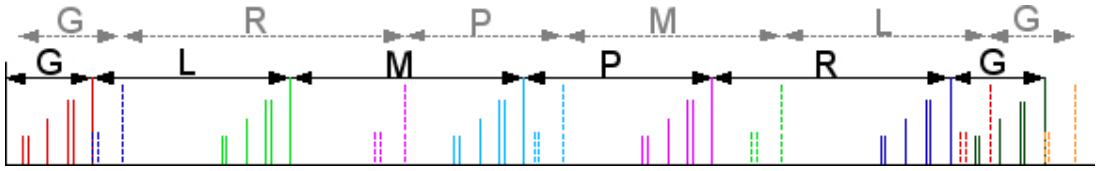


Figure 4.3 – Spectre MS/MS dans lequel les pics en traits continus représentent les ions N-terminaux, et les pics en pointillés représentent les ions C-terminaux. Chaque paquet est distingué à l'aide d'une couleur différente.

## 4.4 Modification des spectres

On peut facilement observer sur la Figure 4.3 que les pics issus d'une même fragmentation, c'est-à-dire les pics faisant partie d'un même paquet $p$, se situent dans deux zones du spectre : une pour les pics N-terminaux, l'autre pour les pics C-terminaux. Or, cette distribution des pics le long du spectre peut nuire à l'alignement d'un spectre expérimental lorsque celui-ci contient une modification -et donc un décalage d'un certain nombre de pics- avec le spectre théorique qui lui correspond.

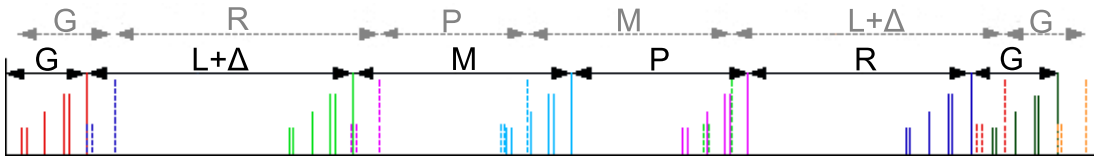


Figure 4.4 – Spectre MS/MS après introduction d'une modification post-traductionnelle sur le second acide aminé. Les traits continus représentent les ions N-terminaux tandis que les pointillés représentent les ions C-terminaux. Chaque paquet est distingué à l'aide d'une couleur différente.

Dans la Figure 4.4 il est possible de visualiser un spectre MS/MS représentant le peptide $G(L+\Delta)MPRG$. Ce peptide correspond au spectre présenté dans la Figure 4.3, mais dans lequel l'acide aminé $L$ a subi une modification post-traductionnelle qui augmente sa masse de $\Delta$ daltons. En observant les spectres produits avec et sans cette modification, il est possible de voir les différences que cela induit sur les paquets, et qui peuvent rendre la comparaison de spectres difficile.

Afin de contourner ce type de problème, nous allons modifier les spectres étudiés, aussi bien le spectre théorique (Section 4.4.1), que le spectre expérimental (Section 4.4.2).

### 4.4.1 Modifications des spectres théoriques

Dans le spectre théorique, nous choisissons de ne plus placer les pics C-terminaux représentant un ion $\beta$ à la position $m(aa_{i+1}...aa_N)+\delta_\beta$ mais à la position $m(P)-(m(aa_{i+1}...aa_N)+\delta_\beta)$ ce qui peut aussi s'écrire $m(aa_1...aa_i)-\delta_\beta$. Nous obtenons donc pour un ion N-terminal $\alpha$ et son ion C-terminal complémentaire $\beta$, des pics situés à une masse $m(aa_1...aa_i)+\delta_\alpha$ pour l'ion N-terminal et à une masse $m(aa_1...aa_i)-\delta_\beta$ pour l'ion C-terminal. Si l'on compare ces deux masses, nous voyons que la différence ne réside que dans les $\delta$, c'est à dire des valeurs constantes ne dépendant que du type d'ion évalué. Un spectre théorique construit de cette manière, c'est-à-dire en donnant à un pic représentant un ion C-terminal $\beta$ une masse de $m(P)-(m(aa_{i+1}...aa_N)+\delta_\beta)$, sera nommé **spectre théorique symétrique**, et noté $SS_t$.

Une constatation importante peut être faite dans un spectre théorique symétrique. Si nous prenons un paquet $p_i$ de ce spectre, où :

- $\alpha_i$ est un pic représentant un ion N-terminal $\alpha$ de $p_i$ et
- $\beta_i$ est un pic représentant un ion C-terminal $\beta$ de $p_i$,

alors toutes les positions des pics composant $p_i$ s'écrivent sous la forme :

- $m(aa_1...aa_i)+\delta_\alpha$ si il s'agit d'un pic représentant un ion N-terminal, ou alors
- $m(aa_1...aa_i)-\delta_\beta$ si il s'agit d'un pic représentant un ion C-terminal.

La symétrie permet ici d'exprimer toutes les positions de pics en fonction de $m(aa_1...aa_i)$ et des $\delta$. Dès lors, dans un spectre théorique symétrique, chaque paquet $p_i$ possède un point, nommé **point de référence** et noté $R_{p_i}$, qui va être utilisé pour définir la position du paquet, c'est-à-dire $R_{p_i}=m(aa_1...aa_i)$.

Ainsi, dans $p_i$, la distance séparant le pic $\alpha_i$ (resp. $\beta_i$) du point $R_{p_i}$ est égale à $+\delta_\alpha$ (resp. $-\delta_\beta$). Comme la valeur $\delta_\alpha$ (resp. $\delta_\beta$) ne dépend que du type d'ion $\alpha$ (resp. $\beta$), et que les ions

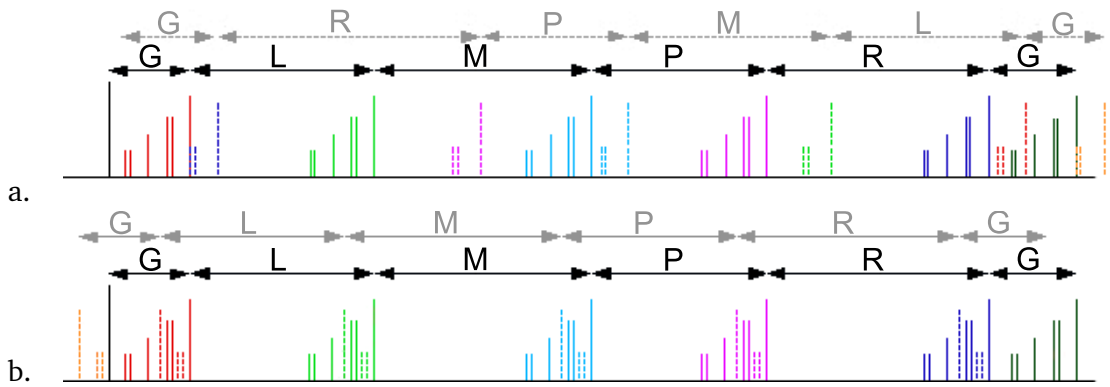


Figure 4.5 – En a. un spectre théorique représentant le peptide GLMPRG. En b. le spectre théorique symétrique du même peptide. Les couleurs utilisées permettent de distinguer les différents paquets au sein des spectres.

rencontrés sont les mêmes dans chacun des paquets (à l'exception des paquets $p_0$ et $p_N$, voir Section 4.3.2), page 48, seule la position adoptée par le paquet changera d'un paquet à l'autre, et non son contenu. La Figure 4.5 montre un spectre théorique avant et après sa transformation pour la prise en compte de la symétrie.

En conséquence de l'application de la symétrie dans le spectre théorique, le paquet devient un modèle des pics attendus à chacun des sites de fragmentation. Un spectre théorique symétrique peut désormais être représenté par un ensemble de paquets, tous identiques et positionnés de manière à délimiter tous les acides aminés. Le contenu d'un modèle de paquet devra être défini pour représenter les ions désirés dans le spectre théorique. La Figure 4.6 représente le modèle de paquet que nous avons défini pour la comparaison à des spectres expérimentaux obtenus sur des appareils de type `QTOF`. Ce modèle, défini comme expliqué en Section 4.4.1.1, page 51, contient les pics représentant les ions : $a$, $a^*$, $a^\circ$, $b$, $b^*$, $b^\circ$, $y$, $y^*$ et $y_\circ$. L'intensité de chacun de ces pics est fonction de la probabilité de présence du pic dans le spectre. En effet, plus un pic du modèle est intense, plus il a de chance d'être observé dans le spectre expérimental.

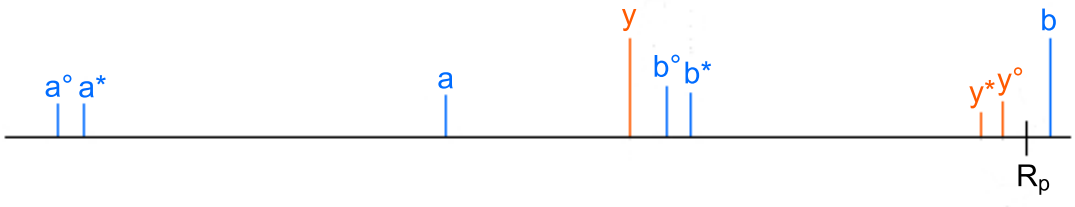


Figure 4.6 – Modèle de paquet que nous avons développé pour les appareils de type `QTOF`.

Pour construire un spectre théorique symétrique de cette manière, il est uniquement nécessaire de connaître le peptide, la masse de chacun des acides aminés et de disposer d'un modèle de paquet. Ensuite, il suffit de positionner un paquet sur chaque site de fragmentation possible, c'est à dire en $m(aa_1...aa_i)\ \forall i \in [1; N]$.

La génération des spectres théoriques correspondant aux différents peptides d'une banque est donc aisée. Le temps nécessaire pour cette génération reste cependant dépendant de la taille de la banque, mais on remarquera que cette génération peut être réalisée une fois pour toutes, donc sous la forme d'un prétraitement.

Un spectre théorique défini de la sorte peut facilement être modifié pour refléter une modification de la séquence qu'il représente. Par exemple, si on désire augmenter la masse du $i$-ème acide aminé de $\Delta_i$, il suffira de translater tous les paquets à droite de cet acide aminé (c'est-à-dire les paquets $p_{i+1}$ à $p_{N+1}$) de $\Delta_i$.

#### 4.4.1.1 Création d'un modèle de paquet

Pour créer le modèle de paquet propre aux appareils `QTOF`, nous nous sommes appuyés sur des études statistiques des fragments présents lors de la fragmentation des peptides avec ce type d'appareil [DAC+99, HHS03, FP05]. Ces études nous ont permis d'identifier les fragments les plus fréquemment observés, que nous avons considérés dans notre modèle de paquet. Nous avons choisi de ne conserver dans notre modèle que les fragments ayant une probabilité strictement supérieure à $15\%$ d'être observés dans le spectre. Les fragments sélectionnés pour notre modèle sont visibles dans la Table 4.1. Lors de la création du modèle, la probabilité d'apparition des pics est utilisée en guise d'intensité. Ainsi, plus un pic du modèle est intense, plus il a de chance d'être observé dans les données expérimentales.

| **Nom du fragment** | **Notation du fragment** | **Valeur du $\delta$** | **Probabilité** |
|---|---|---|---|
| ion y | $y$ | +19 | 0,87 |
| ion y-$NH_3$ | $y^*$ | +2 | 0,24 |
| ion y-$H_2O$ | $y^\circ$ | +1 | 0,26 |
| ion b | $b$ | +1 | 0,83 |
| ion b-$NH_3$ | $b^*$ | -16 | 0,36 |
| ion b-$H_2O$ | $b^\circ$ | -17 | 0,39 |
| ion a | $a$ | -27 | 0,34 |
| ion a-$NH_3$ | $a^*$ | -44 | 0,20 |
| ion a-$H_2O$ | $a^\circ$ | -45 | 0,17 |

Table 4.1 – Liste des fragments conservés dans le modèle de paquet dédié aux appareils `QTOF`.

### 4.4.2 Modifications des spectres expérimentaux

Nous venons de voir que nous avons déplacé certains pics dans le spectre théorique. Il est également nécessaire d'opérer un déplacement de même nature dans les spectres expérimentaux si l'on veut que les spectres expérimentaux et théoriques restent comparables. Mais il est plus compliqué de prendre en compte ces notions dans un spectre expérimental que dans un spectre théorique. En effet, dans un spectre expérimental, différencier un pic N-terminal d'un pic C-terminal est une tâche ardue [YPO+05]. Comme il n'est pas possible de savoir avec certitude quel ion est représenté par un pic, ni même si un pic représente un ion N-terminal ou un

ion C-terminal, il n'est pas possible de modifier ce spectre de la même manière que le spectre théorique. Nous allons donc devoir, pour chaque pic, considérer les deux hypothèses suivantes :

1. le pic représente un ion N-terminal : dans ce cas, sa position n'est pas modifiée, ou
2. le pic représente un ion C-terminal : ce pic, présent à la position $m$, doit alors être déplacé à la position $m(P) - m$. La masse $m(P)$ est ici connue, puisqu'il s'agit de la masse du précurseur, fournie par le spectromètre de masse.

Pour chaque pic, nous allons donc générer un nouveau pic complémentaire correspondant à l'hypothèse que celui-ci représente un ion C-terminal. Le spectre expérimental traité de la sorte sera nommé **spectre expérimental symétrique**, et noté $SS_e$. Le pic répondant à l'hypothèse 1. ci-dessus sera nommé **pic d'origine**, tandis que le pic généré pour répondre à l'hypothèse 2. sera qualifié de **complémentaire**. Nous dirons aussi que le pic complémentaire est le **symétrique** du pic d'origine (et inversement).

La Figure 4.7 représente un spectre expérimental symétrique. La partie supérieure à l'axe horizontal répond à l'hypothèse 1. et correspond aux pics fournis en sortie du spectromètre de masse. La partie inférieure à l'axe horizontal contient les pics complémentaires créés et utilisés pour traiter l'hypothèse 2.

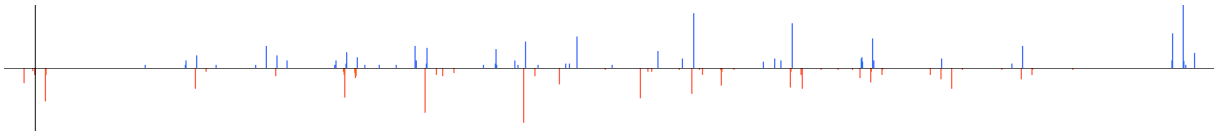

Figure 4.7 – Création d'un spectre expérimental symétrique. Les pics au-dessus de l'axe horizontal sont les pics d'origine du spectre expérimental, les pics sous l'axe horizontal sont les pics complémentaires ajoutés.

### 4.4.3 Particularités de l'alignement des spectres symétriques

Le spectre théorique symétrique $SS_t$ et le spectre expérimental symétrique $SS_e$ obtenus après transformation sont désormais comparables. Il est donc possible de rechercher le meilleur alignement entre ces deux spectres. Cependant, nous pouvons noter deux particularités qui distinguent cet alignement de l'alignement "classique" entre un spectre théorique et un spectre expérimental non transformés.

#### 4.4.3.1 Particularité liée à la duplication des pics

La génération du spectre expérimental symétrique a multiplié le nombre de pics par deux, multiplication liée à l'ajout des pics complémentaires. Sans précaution particulière, les pics complémentaires risquent de nuire aux résultats en se comportant comme du bruit. Nous devons donc prendre soin de vérifier la concordance des pics durant l'alignement.

Un pic C-terminal d'un des paquets du spectre théorique ne pourra être aligné que sur un pic complémentaire du spectre expérimental. En effet, comme nous l'avons défini dans la

Section 4.4.2, les pics complémentaires du spectre expérimental sont les pics supposés être C-terminaux.

Par exemple, les pics en trait continu (resp. en pointillé) dans la Figure 4.5 b de la page 50. ne seront alignés qu'avec les pics situés au-dessus (resp. au-dessous) de l'axe horizontal dans la Figure 4.7.

#### 4.4.3.2 Particularité de l'alignement de paquets : les positions possibles

D'une manière générale, la méthode d'alignement de deux spectres consiste à essayer de faire correspondre au mieux chacun des pics du spectre expérimental avec un pic du spectre théorique. Cependant, nous avons vu que les pics du spectre théorique symétrique sont regroupés en paquets. Nous appellerons **ensemble des positions possibles d'un spectre expérimental symétrique** l'ensemble des positions de ce spectre pour lesquelles il est utile d'évaluer le score de l'alignement avec un paquet, c'est-à-dire des positions permettant d'aligner des pics.

Le nombre de positions possibles a donc pour borne supérieure le nombre de pics du spectre expérimental. Comme nous le verrons dans le Chapitre 6, il est également possible d'augmenter les contraintes d'alignement (par exemple imposer qu'au moins deux pics correspondent entre un paquet et un spectre expérimental symétrique) et en conséquence de limiter le nombre de positions possibles.

Lorsqu'un paquet est aligné sur une position possible, son point de référence ($R_p$) est placé à la position représentée par cette position possible.

## 4.5 Algorithme d'alignement de deux spectres

Dans cette partie, nous présentons notre nouvel algorithme, basé sur une approche par programmation dynamique, visant à comparer efficacement un spectre théorique symétrique à un spectre expérimental symétrique. Comme il en a été fait mention précédemment, cette méthode doit tolérer un nombre important de modifications entre les deux spectres.

### 4.5.1 Principe de la programmation dynamique

Le terme de *programmation dynamique* a été introduit par Richard Bellman dans les années 40, puis formalisé sous sa forme actuellement connue en 1952 dans [Bel52].

La **programmation dynamique** est une méthode de programmation visant à résoudre des problèmes complexes, en les découpant en sous-problèmes plus simples de manière récursive. Elle a pour but de résoudre des problèmes d'optimisation en recherchant une solution associée à une valeur optimale. Pour obtenir cette valeur optimale, l'algorithme utilise les valeurs des sous-problèmes précédemment calculés.

Notre problématique d'alignement de deux spectres est bien adaptée à une résolution par programmation dynamique, car elle respecte un ensemble de propriétés caractéristiques :

– Le problème peut se diviser en étapes, chacune des étapes nécessitant une prise de décision.
  Chaque étape correspond à un score d'alignement d'un paquet avec une position possible. À chaque étape, il faut décider si oui ou non le paquet doit être aligné sur la position possible, en effectuant une modification si nécessaire.

La qualité d'un alignement peut s'évaluer par un score qui est lié à la concordance des pics entre les paquets et les pics qui entourent une position possible du spectre expérimental. Une solution optimale d'alignement sera associée au score d'alignement maximal qui correspond à la somme des scores d'alignement de chacun des paquets du spectre théorique.

- Chacune des étapes est associée à un ou plusieurs états.
  L'étape courante a comme état le score du meilleur alignement pour tous les paquets et positions possibles déjà traités.

- La décision prise à une étape va transformer un état en un nouvel état à l'étape suivante. Décider d'aligner ou non un paquet sur une position possible va modifier l'état, c'est-à-dire le score de l'alignement, ce score pouvant augmenter ou diminuer.

- Étant donné l(es) état(s) courant(s), la décision optimale pour définir chacun des états restants ne dépend pas des états précédents, mais uniquement de celui(ceux) de l'état courant. En effet, pour compléter un alignement avec l'alignement de nouveaux paquets sur de nouvelles positions possibles, il n'est pas nécessaire de savoir quels sont les paquets précédemment alignés, mais uniquement de connaître le score de l'alignement à l'étape courante.

Ces différentes caractéristiques nous permettent de valider le fait que la programmation dynamique est une méthode parfaitement appropriée pour résoudre notre problème, et permettent de définir la récurrence de notre problème.

### 4.5.2 Les paramètres de l'algorithme d'alignement

Les différents paramètres fournis en entrée de notre algorithme d'alignement sont :
- Le spectre théorique symétrique $SS_t$ qui est défini par sa liste de positions des paquets.
- Le spectre expérimental symétrique $SS_e$ qui est défini par la liste de ses positions possibles.
- Un entier $K$ qui détermine le nombre maximal de modifications tolérées durant l'alignement. Ce paramètre $K$ est primordial dans la méthode. S'il vaut 0, alors l'algorithme d'alignement ne pourra effectuer qu'une modification de type "recalibrage", c'est-à-dire une modification en tout début de spectre visant à corriger un potentiel mauvais calibrage du spectromètre de masse.

### 4.5.3 Le résultat d'un alignement

Notre méthode fournit en sortie différents éléments, tout d'abord : (1) le meilleur score d'alignement trouvé entre les deux spectres. (2) la liste des modifications appliquées lors de l'alignement, cette liste pouvant bien évidement être vide si aucune modification n'a été nécessaire afin d'obtenir le meilleur alignement. Chacune des modifications se traduit par une position dans le spectre (en terme de $m/z$), la masse de l'acide aminé présent avant modification ainsi que sa masse après modification. Il est donc ainsi possible de connaître chacune des modifications et de reconstituer le spectre théorique correspondant au spectre expérimental tel qu'il a été réellement observé.

### 4.5.4 Le score mesurant la similarité entre deux spectres

Différentes méthodes de calcul de la similarité entre deux spectres ont été présentées dans le Chapitre 3. Malheureusement, aucune de ces méthodes connues ne peut être directement appliquée à notre situation. Le score de similarité entre les deux spectres doit tenir compte de manière additive de la similarité des paquets alignés sur les positions possibles, afin de respecter la contrainte imposée par l'usage de la programmation dynamique. Il doit également pénaliser l'introduction des modifications pour que celles-ci ne soient pas trop nombreuses. En fonction des différents paramètres de l'algorithme, l'évaluation empirique de la pertinence de différentes méthodes de calcul de scores sera présentée dans le Chapitre 6, page 69. Dans un premier temps, nous supposerons donc que la fonction de score est connue.

### 4.5.5 L'algorithme d'alignement PacketSpectralAlignment

Notre algorithme d'alignement, **PacketSpectralAlignment**, construit deux matrices, $M$ et $D$, toutes deux de dimension 3. La première dimension de ces matrices représente les paquets du spectre théorique symétrique, la seconde dimension représente les positions possibles du spectre expérimental symétrique. Les deux premières dimensions sont associées à un alignement sans modification. La troisième dimension permet la recherche du meilleur alignement après l'introduction d'au plus $K$ modifications.

- $M(i,j,k)$ représente le meilleur score trouvé lors de l'alignement des paquets $p_1$ à $p_i$ sur les positions possibles $pp_1$ à $pp_j$ avec au plus $k$ modifications, avec $k \leq K$.
- $D(i,j,k)$ représente le meilleur score trouvé pour l'alignement des paquets $p_1$ à $p_i$ sur les positions possibles $pp_1$ à $pp_j$ tel que le paquet $p_i$ soit obligatoirement aligné sur la position possible $pp_j$, le tout avec au plus $k$ modifications, avec $k \leq K$.

De plus nous définissons que :

- $(i',j') < (i,j)$ si et seulement si $i' < i$ et $j' < j$
- $(i',j') \leq (i,j)$ si et seulement si $i' \leq i$ et $j' \leq j$
- $(i',j')$ est précurseur de $(i,j)$ si $(i',j') < (i,j)$ et si $Rp_i - Rp_{i'} = pp_j - pp_{j'}$ (pour rappel, $Rp_i$ correspond au point de référence du paquet $p_i$)
- $(0,0)$ est toujours précurseur de $(i,j)$, ceci pour des raisons d'initialisation
- $D(0,0,k) = 0$ pour des raisons d'initialisation
- score($p_i$, $pp_j$) donne le score résultant de l'alignement des pics du paquet $p_i$ de $SS_t$ sur les pics de $SS_e$ quand le point de référence $Rp_i$ est aligné sur la position possible $pp_j$ de $SS_e$.

Nous pouvons ensuite écrire les formules de récurrence suivantes, lesquelles seront utilisées dans l'algorithme de programmation dynamique :

$$D(i,j,k) = \max_{(i',j')<(i,j)} \begin{cases} D(i',j',k) + \text{score}(p_i, pp_j) & \text{si } (i',j') \text{ est un précurseur de } (i,j) \\ D(i',j',k-1) + \text{score}(p_i, pp_j) & \text{sinon} \end{cases} \tag{4.2}$$

$$M(i,j,k) = \max_{(i',j')\leq(i,j)} D(i',j',k) \tag{4.3}$$

Définissons Dprécurseur$(i, j, k) = \max\{D(i', j', k) \mid (i', j')$ précurseur de $(i, j)\}$.

Et la fonction récupérerPP($j$, `PacketSize`) = $j'$ tel que $pp_{j'} = \max\{pp_{j''} \mid pp_{j''} \leq pp_j -$ `PacketSize` }, où `PacketSize` représente la taille d'un paquet.

Nous pouvons alors réécrire les récurrences (4.2) et (4.3) sous la forme suivante :

$$D(i,j,k) = \max \begin{cases} \text{Dprécurseur}(i,j,k) + \text{score}(p_i, pp_j) \\ M(i-1, \text{récupérerPP}(j, \texttt{PacketSize}), k-1) + \text{score}(p_i, pp_j) \end{cases} \tag{4.4}$$

$$M(i,j,k) = \max \begin{cases} D(i,j,k) \\ M(i-1,j,k) \\ M(i,j-1,k) \end{cases} \tag{4.5}$$

L'algorithme PacketSpectralAlignment (Algorithme 1) va donc remplir les valeurs $M(i, j, k)$ et $D(i, j, k)$ pour (1) chaque nombre possible de modifications $k$ jusqu'à un maximum de $K$, (2) chaque position possible $pp_j$ de $SS_e$ et (3) chaque paquet $p_i$ de $SS_t$.

La matrice $D$ est ensuite mise à jour (Algorithme 1, ligne 4) en choisissant la meilleure des possibilités suivantes :

1. Nous pouvons améliorer un alignement déjà existant (donné par Dprécurseur) sans appliquer de nouvelles modifications.
2. Nous devons appliquer une modification pour aligner $p_i$ avec $pp_j$ : dans ce cas, nous devons utiliser le meilleur alignement trouvé dans la matrice $M$ avec $k-1$ modifications, que l'on peut prolonger par l'alignement de $p_i$ avec $pp_j$. Le score de cet alignement se trouve en $M(i-1,$ récupérerPP($j$, `PacketSize`)$, k-1)$. Le premier paramètre est $i-1$ car le paquet $p_i$ ne doit pas en faire partie (sinon il se retrouverait aligné 2 fois). Le deuxième paramètre est récupérerPP($j$, `PacketSize`) car il faut s'assurer que lorsque nous appliquerons une modification, l'alignement restera réalisable. La constante `PacketSize` va nous permettre de nous assurer que, même après modification, l'espace entre deux paquets soit "réaliste" (pas de chevauchement de paquets, une distance avec le paquet précédent qui correspond à un acide aminé plausible potentiellement porteur d'une modification post-traductionnelle réelle). On notera aussi que le score est ajusté grâce à la valeur `Pénalité` qui va pénaliser l'application d'une modification dans l'alignement.

La matrice $M$ peut être mise à jour (Algorithme 1, ligne 5) en sélectionnant le meilleur alignement trouvé jusqu'à ce point (que ce soit en alignant le paquet $p_i$ sur la position possible $pp_j$ ou non).

Ensuite, des étapes visant à récupérer le meilleur alignement dans ces matrices vont se succéder. Dans un premier temps, il s'agit de rechercher pour quel nombre $k$ de modifications le meilleur alignement a pu être obtenu. En effet, une modification entraînant une pénalité dans le score (via la constante `Pénalité` dans l'Algorithme 1, ligne 4), le meilleur score ne sera pas nécessairement obtenu pour le plus grand nombre de modifications. Ce nombre de modifications permet d'accéder au meilleur score, et il est mémorisé dans la variable $meilleur\_k$

**Algorithm 1** PacketSpectralAlignment : alignement

**Entrée :**
Ensemble des Paquets de $SS_t$ : $\{p_i | \forall i \in [1; N+1]\}$,
Ensemble de Positions Possibles de $SS_e$ : $\{pp_j | \forall j \in [1\,;\, Q]\}$ et
Entier $K$

**Sortie :**
Réel $score\_final$ et
Ensemble de modifications $modifications$

1: **pour** $k$ de 0 à $K$ **faire**
2: $D(0,0,k) = 0$
3: **fin pour**
4: **pour** $k$ de 0 à $K$ **faire**
5: **pour** $j$ de 1 à $Q$ **faire**
6: **pour** $i$ de 1 à $N+1$ **faire**
7: $D(i,j,k) = \max$(Dprécurseur$(i,j,k)$ + score$(p_i, pp_j)$,
$M(i-1,$ récupérerPP($j$, `PacketSize`) $, k-1)$ + score$(p_i, pp_j)$ + `Pénalité`)
8: $M(i,j,k) = \max(D(i,j,k),\ M(i-1,j,k),\ M(i,j-1,k))$
9: **fin pour**
10: **fin pour**
11: **fin pour**
12: **pour** $k$ de 0 à $K$ **faire**
13: **si** $score\_final < M(N+1,Q,k)$ **alors**
14: $meilleur\_k = k$
15: $score\_final = M(N+1,Q,k)$
16: **fin si**
17: **fin pour**
18: $modifications$ = backtrack( $meilleur\_k$ )

(Algorithme 1, ligne 11), tandis que le meilleur score est mémorisé dans la variable de sortie $score\_final$ (Algorithme 1, ligne 12).

L'Algorithme 1 présente une complexité temporelle de l'ordre de $O(NQK)$ où $N$ est le nombre d'acides aminés du peptide représenté par le spectre théorique (donc $N+1$ paquets), $Q$ le nombre de positions possibles du spectre expérimental et $K$ le nombre maximum autorisé de modifications. La complexité spatiale est, également, de l'ordre de $O(NQK)$.

Pour conclure, notre algorithme effectue une dernière étape dite de **backtrack**. Cette étape, caractérisée ici par l'appel à une fonction **backtrack** (Algorithme 1 ligne 15), consiste à déterminer, une fois le meilleur score obtenu, comment celui-ci a été produit. Cette fonction prend donc en paramètre le point $(i, j, k)$ où se situe la meilleure solution, afin de remonter jusqu'au point de départ dans la matrice. Ce backtrack va ainsi produire en sortie la liste des modifications nécessaires pour obtenir le résultat donné.

Pour rendre le backtrack possible, il est nécessaire de mémoriser pour chaque élément des matrices $D$ et $M$ leur origine, c'est-à-dire les coordonnées de la cellule précédemment utilisée dans l'alignement. Pour cela, nous utilisons deux matrices $M\_origine$ et $D\_origine$, qui vont contenir ces coordonnées. Ces matrices sont remplies durant le déroulement de l'Algorithme 1, mais par soucis de compréhension et de lisibilité, nous avons choisi de ne pas les y faire figurer. Ensuite, en partant du dernier alignement de la matrice $M$ ($M(N, Q-1, K-1)$), pour connaître les modifications, il faut remonter jusqu'à l'élément $M(0, 0, 0)$. Cette remontée va se faire un utilisant principalement la matrice $D\_origine$. Si à un moment il n'est plus possible de progresser dans la remonté en utilisant cette matrice, alors la matrice $M\_origine$ est consultée et va permettre de savoir si il y a eu, ou non, une modification (changement de $k$ dans les coordonnées mémorisées).

L'Algorithme 2 présente cette fonction de backtrack. Dans cet algorithme, la fonction **position_possible**($j$) permet de connaître la position (en terme de masse) de la $j$-ème position possible dans le spectre expérimental. La fonction position_paquet($i$) va quant à elle donner la position (toujours en terme de masse) du $i$-ème paquet dans le spectre théorique. À la ligne 28 de l'Algorithme 2 est créée une modification. Il s'agit d'un ensemble de valeurs la définissant. Ainsi, **Modification**($i$, $taille$, $ancienne\_taille$) définit une modification ayant lieu sur l'acide aminé suivant le $i$-ème paquet du spectre théorique, modifiant sa masse de $ancienne\_taille$ à $taille$. À la ligne 20, les valeurs $-1$ sont les valeurs d'initialisation utilisées dans la matrice $D\_origine$, elles apparaissent lorsqu'il n'est pas possible de remonter plus loin et qu'il faut utiliser les valeurs de la matrice $M\_origine$.

La complexité du backtrack, tel que présenté ici, est de l'ordre de $O(Q)$.

**Algorithm 2** PacketSpectralAlignment : backtrack

**Entrée :**
Entier $N$ nombre d'acides aminés de $SS_t$, $Q$ nombre de positions possibles de $SS_e$ et $K$ nombre de modifications tolérées
Matrices $M$ et $D$ remplie par l'Algorithme 1
Matrices $M\_origine$ et $D\_origine$ contenant pour chaque cellule de $M$ et $D$ les coordonnées de l'élément précédemment utilisé dans l'alignement.

**Sortie :**
Ensemble de modifications $modifications$

1: $score = 0$
2: $nb\_mod = 0$
3: $modifications = \emptyset$
4: **pour** $k$ de 0 à $K$ **faire**
5:  **si** $score < M(N, Q-1, k)$ **alors**
6:  $score = M(N, Q-1, k)$
7:  $nb\_mod = k$
8:  **fin si**
9: **fin pour**
10: $i = N$ ; $j = Q-1$ ; $k = nb\_mod$
11: **tant que** $i \leq 0$ **et** $j \leq 0$ **faire**
12:  $(t_i, t_j, t_k) = M\_origine(i, j, k)$
13:  **si** $t_i = i$ **et** $t_j = j$ **alors**
14:  $(t_i, t_j, t_k) = D\_origine(i, j, k)$
15:  **si** $t_k = k$ **alors**
16:  $t_i = i$ ; $t_j = j$
17:  $(i, j, k) = D\_origine(t_i, t_j, k)$
18:  **sinon**
19:  $t_i = i$ ; $t_j = j$ ; $t_k = k$
20:  $(i, j, k) = D\_origine(t_i, t_j, t_k)$
21:  **si** $i \neq -1$ **et** $j \neq -1$ **alors**
22:  $(tt_i, tt_j, tt_k) = M\_origine(i, j, k)$
23:  **tant que non(** $tt_i = i$ **et** $tt_j = j$ **)** **faire**
24:  $ttt_i = i$ ; $ttt_j = j$
25:  $(i, j, k) = M\_origine(ttt_i, ttt_j, k)$
26:  **fin tant que**
27:  $taille$ = position_possible($t_j$) - position_possible($j$)
28:  $ancienne\_taille$ = position_paquet($i+1$) - position_paquet($i$)
29:  $mod$ = Modification( $i$, $taille$, $ancienne\_taille$)
30:  $modifications = modifications \cup mod$
31:  **fin si**
32:  **fin si**
33:  **sinon**
34:  $t_i = i$ ; $t_j = j$
35:  $(i, j, k) = M\_origine(t_i, t_j, k)$
36:  **fin si**
37: **fin tant que**

# CHAPITRE 5

# Jeux de données et critères d'évaluation

## 5.1 Introduction

L'algorithme PacketSpectralAlignment que nous avons présenté Chapitre 4 se situe au cœur du processus d'identification des protéines, en associant des peptides aux spectres expérimentaux. Pour faire face à des contraintes d'efficacité et de précision, une approche empirique est maintenant nécessaire pour extraire et estimer au mieux les paramètres de fonctionnement de l'algorithme et limiter les traitements aux spectres les plus pertinents.

Dans ce chapitre, nous allons décrire la constitution des jeux de données qui vont permettre l'apprentissage des différents paramètres de l'algorithme, puis nous présenterons les critères qui ont été utilisés pour évaluer son comportement et mesurer la pertinence de notre approche.

## 5.2 Jeux de données

L'ajustement des paramètres de l'algorithme PacketSpectralAlignment nécessite de disposer de jeux de données composés de spectres MS/MS associés avec une grande confiance à des identifications de peptides puis de protéines. L'idéal serait bien sûr de disposer d'une banque de référence internationale de milliers de spectres obtenus à partir de peptides de synthèse dont les séquences seraient parfaitement connues. Malheureusement, une telle banque n'est pas disponible.

Néanmoins, de nombreux entrepôts stockant des spectres de masse MS ou MS/MS ont été mis en place ces dernières années, parmi lesquels nous pouvons citer PeptideAtlas [DLA08], PRIDE [VCR+09] ou le très prometteur Peptidome [SBE09, JBA+10] du NCBI. La plupart de ces entrepôts ont pour objectif de faciliter le partage de données lors de collaborations dans des projets de grande envergure, ou de permettre une meilleure diffusion des données lors de la publication de résultats. Malheureusement, les jeux de données que l'on pourrait extraire de ces entrepôts ne répondent pas totalement à nos besoins. En effet, la description des métadonnées qui accompagnent les spectres n'est souvent pas suffisante dans notre contexte : par exemple, l'identification correcte des spectres manque parfois, d'autre part, il est rarement fait mention du type de spectromètre employé. De plus, l'identification associée aux spectres ne peut pas être qualifiée de sûre. Nous présentons ci-après les différents jeux de données de spectres que nous avons sélectionnés compte tenu de nos contraintes, ainsi que les différentes banques de protéines que nous avons élaborées pour créer un environnement d'apprentissage et d'évaluation.

### 5.2.1 Jeux de données de spectres

#### 5.2.1.1 Jeu de données de l'ISB

L'institut des systèmes biologiques de Seattle (*Institute for Systems Biology* : ISB) a créé spécifiquement des jeux de données pour qu'ils servent à évaluer les différentes méthodes de traitement de spectres MS/MS. Ces jeux de données [KEJ+08] sont d'un intérêt majeur car le produit analysé est un mélange de 18 protéines connues et disponibles dans les banques de protéines (la liste de ces protéines est présentée dans le Tableau 5.1). De plus, un soin particulier a été accordé au protocole des expériences pour garantir une excellente reproductibilité des résultats.

| **Protéine** | **Organisme** | **Identifiant Swiss-Prot** | **Masse (kDa)** |
|---|---|---|---|
| Actin, aortic smooth muscle | Bovine | P62739 | 42.0 |
| Alkaline phosphotase | *E. coli* | P00634 | 49.4 |
| Alpha-amylase | *B. licheniformis* | P06278 | 58.5 |
| Alpha-lactalbumin | Bovine | P00711 | 16.2 |
| Beta-casein | Bovine | P02666 | 25.1 |
| Beta-galactosidase | *E. coli* | P00722 | 116.5 |
| Beta-lactoglobulin | Bovine | P02754 | 19.9 |
| Carbonic anhydrase 2 | Bovine | P00921 | 29.1 |
| Catalase | Bovine | P00432 | 59.9 |
| Cytochrome c | Bovine | P62984 | 11.7 |
| Glyceraldehyde-3-phosphate dehydrogenase | Rabbit | P46406 | 35.8 |
| Glycogen phosphorylase, muscle form | Rabbit | P00489 | 97.3 |
| Mannose-6-phosphate isomerase | *E. coli* | P00946 | 42.9 |
| Myoglobin | Horse | P68082 | 17.1 |
| Myosin light chain 1, skeletal muscle isoform | Rabbit | P02602 | 20.9 |
| Ovalbumin | Chicken | P01012 | 42.9 |
| Serotransferrin | Bovine | Q29443 | 77.8 |
| Serum albumin | Bovine | P02769 | 69.3 |

Table 5.1 – Liste des protéines contenues dans le mélange analysé pour produire les données de l'ISB.

Les jeux de données se présentent sous la forme de 4 ensembles de spectres, correspondant à l'analyse de 4 échantillons appelés *mix 1*, *mix 2*, *mix 3* et *mix 4*. Les échantillons contiennent le même mélange de protéines, mais ils correspondent à des variantes de protocoles expérimentaux décrits dans [KEJ+08]. Les quatre *mix* ont été analysés par huit modèles différents de spectromètres de masse en mode MS/MS, ces modèles couvrant les appareils les plus usuels au moment de l'étude en 2006-2007. Un nombre important de répétitions a été effectué pour chacun des appareils. En général, les *mix* ont été analysés 10 fois sur chacun des appareils, même si toutes les combinaisons *mix*/appareil n'ont pas été effectuées. Au final, le jeu de données de l'ISB est constitué de 150 ensembles de spectres MS/MS, chaque ensemble étant associé à un *mix* et un appareil particulier.

Les spectres MS/MS sont disponibles à la fois en format brut (format RAW dépendant du fabricant du spectromètre) et en format mzXML proposé comme un format standard en spectrométrie de masse [PEH+04].

A l'issue de chaque analyse MS/MS, les spectres ont été associés aux peptides des protéines du mélange à l'aide de l'outil Sequest, puis validés par le logiciel PeptideProphet. Ainsi, lorsqu'une association a été possible entre un spectre et un peptide, ce dernier est disponible dans un fichier de résultat avec son score de *p-value* qui permet d'évaluer la fiabilité de l'association.

Dans le jeu de l'ISB, nous avons sélectionné arbitrairement le second échantillon *mix* 2 analysé sur le spectromètre de masse de type `QTOF` annoté QTOF 1. Ce type d'appareil est celui qui a été utilisé à l'INRA de Nantes durant cette thèse.

Le Tableau 5.2 fournit des indications sur les spectres constituant le second jeu de données du *mix* 2. Différents prétraitements ont été appliqués sur les spectres bruts de ce jeu de données afin de les déconvolutionner et d'en retirer les isotopes et le bruit en utilisant l'outil Mascot Distiller [Mat] avec ses paramètres par défaut. Ces différents prétraitements ont été précédemment évoqués en Section 2.3.4.3, page 22. Mascot Distiller est un outil commercial distribué par Matrix Science qui utilise une approche similaire à celles détaillées dans [BHL99, GMG+99]. Cet outil va donc permettre de nettoyer les données et de combiner les spectres représentant un même peptide pour obtenir un seul spectre représentatif. De plus, seuls les spectres pour lesquels un peptide associé est disponible dans les résultats de Sequest sont conservés dans le jeu de données final après traitement. Comme il est montré dans ce tableau, les prétraitements réduisent le nombre de spectres de près de 3000 spectres bruts à 1063 spectres prétraités. Nous nommerons ce jeu de données **spectres_ISB** .

| | |
|---|---|
| Nombre de spectres bruts | 2989 |
| Nombre de spectres après les prétaitements et ayant une identification associée | 1063 |
| Nombre moyen de pics par spectre | 595 |
| Nombre minimal de pics dans un spectre | 12 |
| Nombre maximal de pics dans un spectre | 1596 |
| Longueur moyenne d'un spectre (en daltons) | 660 |
| Longueur minimale d'un spectre (en daltons) | 300 |
| Longueur maximale d'un spectre (en daltons) | 1369 |

Table 5.2 – Description du jeu de données **spectres_ISB**.

#### 5.2.1.2 Brachypodium

Notre second jeu de données résulte de multiples analyses d'une protéine provenant de la plante Brachypodium distachyon (Figure 5.1). L'intérêt majeur de ce jeu de données est qu'il s'agit d'un organisme représentatif de ceux étudiés par le centre INRA de Nantes, à savoir un végétal modèle pour le blé, et qu'il a été obtenu sur le même type de spectromètre de masse

que les données spectres_ISB mais avec un appareil et des réglages différents. Il permettra ainsi de vérifier qu'il n'y a pas eu de sur-apprentissage de nos paramètres sur les données spectres_ISB.

Figure 5.1 – Brachypodium distachyon. *Source : Dawson, J.E. and Hatch, S.L.*.

Les données de Brachypodium correspondent à 8 jeux de spectres expérimentaux issus de l'analyse de 8 spots (D1 à D8) d'un gel d'électrophorèse 2D, tous ces spots contenant a priori une unique protéine : une globuline nommée Bradi1g13040.1 [LPB+10]. Pour l'identification de ces spectres, nous nous attendons donc à retrouver des peptides de cette protéine, mais nous ne disposons cependant pas d'une liste qui associe les différents spectres avec les peptides de la protéine. Le Tableau 5.3 donne le nombre de spectres présents dans les 8 jeux de spectres des analyses des spots D1 à D8. Nous nommerons ces 8 jeux de spectres **bradi_x** où x est le nom du spot concerné (le jeu de spectres du spot D2 sera donc nommé bradi_D2). Nous nommerons **bradi_total** le jeux de spectres combinant les 8 spots de brachypodium.

| | **D1** | **D2** | **D3** | **D4** | **D5** | **D6** | **D7** | **D8** | **Moyenne** |
|---|---|---|---|---|---|---|---|---|---|
| **Nombre de spectres** | 290 | 248 | 239 | 121 | 145 | 356 | 142 | 90 | 204 |

Table 5.3 – Nombre de spectres présents dans chacun des jeux de spectres issus de l'analyse de Bradi1g13040.1.

Nous avons pu analyser les 8 jeux de spectres une première fois en utilisant le logiciel Mascot, ce qui a permis de confirmer la protéine à retrouver avec notre méthode. Les spectres issus des différentes analyses comportent en effet de nombreux peptides non modifiés. De plus nous nous attendons à rencontrer des modifications post-traductionnelles chimiques spécifiques du traitement des échantillons lors de l'analyse des spectres.

Utiliser ce jeu de données va permettre de valider le comportement de notre méthode sur un jeu de données assez différent de celui de l'ISB, montrant par la même occasion que nos réglages et nos choix ne sont pas spécifiques aux données de l'ISB.

### 5.2.2 Banques de données

#### 5.2.2.1 Banque de données pour les spectres spectres_ISB

Nous utiliserons généralement deux banques différentes lorsque nous comparerons les spectres du jeu spectres_ISB à des peptides. La première contiendra uniquement les 18 protéines supposées être dans l'échantillon. Nous appellerons cette banque **18mix**. Puis, nous serons amenés à travailler sur une banque de taille plus conséquente, constituée des 18 protéines de l'échantillon auxquelles 1700 protéines choisies aléatoirement dans la banque du riz ont été ajoutées. Cette variante de la banque sera nommée **18mix_rice1700**. Le Tableau 5.4 précise le nombre de protéines et le nombre de peptides de chacune de ces deux banques lorsque nous ne conservons que les peptides compris entre 600 et 3000 daltons, ce qui correspond à la tolérance habituelle d'un spectromètre de type `QTOF`.

| | 18mix | 18mix_rice1700 |
|---|---|---|
| Nombre de protéines | 18 | 1718 |
| Nombre de peptides | 454 | 28859 |

Table 5.4 – Taille des banques utilisées pour comparer les spectres spectres_ISB

#### 5.2.2.2 Banque de données pour les spectres de Brachypodium

Brachypodium est une plante récemment séquencée [VGM+10], nous disposons donc de sa banque de protéines. Nous appellerons cette banque **Bradi**. Par soucis de temps d'exécution lors de nos différents tests, nous avons créé une sous banque de brachypodium ne contenant que le premier chromosome. Nous appellerons cette banque **Bradi_1g**. Le Tableau 5.5 précise le nombre de protéines et le nombre de peptides de chacune de ces deux banques lorsque nous ne conservons que les peptides compris entre 600 et 3000 daltons, ce qui correspond à la tolérance habituelle d'un spectromètre de type `QTOF`.

| | Bradi_1g | Bradi |
|---|---|---|
| Nombre de protéines | 9444 | 32255 |
| Nombre de peptides | 166178 | 545052 |

Table 5.5 – Taille des banques Bradi et Bradi_1g.

#### 5.2.2.3 Ajout de modifications dans un jeu de données

L'obtention d'un jeu de données expérimentales comportant des modifications parfaitement identifiées est très difficile. La manière la plus simple de l'obtenir est d'introduire des modifications au niveau des protéines de référence et non au niveau des spectres.

Nous avons donc choisi de modifier la banque de protéines associée au jeu de spectre spectres_ISB en introduisant des modifications sur chacune des protéines du mélange. De cette

manière, il existera des différences entre les peptides réellement représentés par les spectres et les peptides situés dans la banque de protéines. Il existe de très nombreuses manières de modifier la banque de protéines, mais pour nous approcher d'une situation d'analyse de protéines issues d'organismes non séquencés, nous avons choisi de simuler les mutations entre espèces au cours du temps à l'aide de matrices PAM.

**Modification d'une banque à l'aide d'une matrice PAM.** Une matrice PAM est une matrice contenant, pour chaque acide aminé, la probabilité qu'il a de muter en un autre au cours de l'évolution [DSO78]. Par définition, une matrice PAM (ou PAM1) est définie de sorte à induire une modification pour cent acides aminés. Différentes matrices PAM peuvent être extrapolées pour des taux de mutation plus importants. Par exemple, une matrice PAM40 correspond à 40 applications successives d'une matrice PAM1, ce qui aura pour conséquence d'augmenter le nombre moyen de mutations.

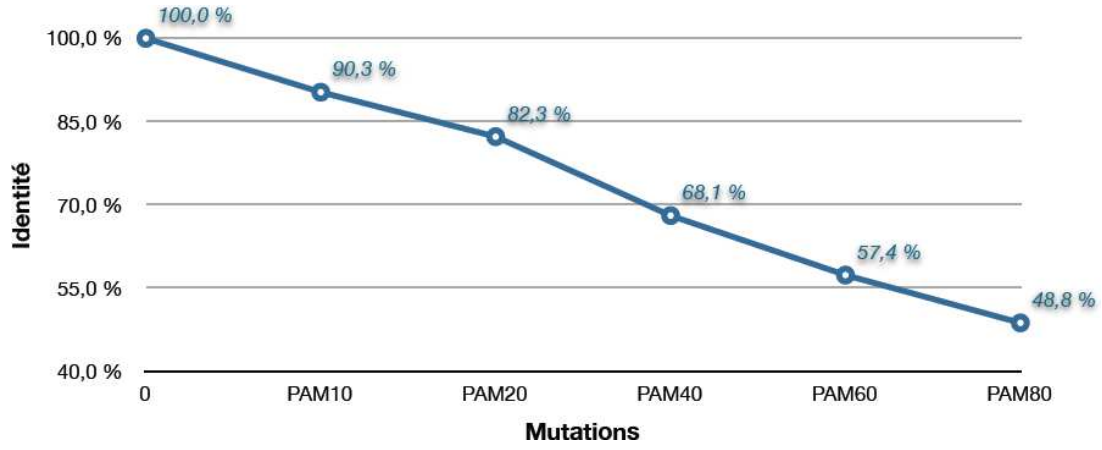


Figure 5.2 – Mesure de l'identité entre les séquences peptidiques non modifiées avec leur version modifiée en utilisant différentes matrices PAM. Les mesures ont été réalisées sur la banque 18mix_rice1700.

En choisissant avec soin la matrice PAM à utiliser pour modifier la banque, il est possible d'approximer la distance phylogénétique séparant la version non modifiée de la banque de sa version modifiée. La Figure 5.2 présente le résultat de la modification des séquences d'une banque par différentes matrices PAM en comparant les séquences non modifiées avec les séquences modifiées. Une valeur de $100\%$ signifie que les séquences sont parfaitement identiques, tandis que $0\%$ signifie que les séquences n'ont aucun acide aminé en commun. Nous pouvons aussi constater que la mutation de la banque par une matrice PAM80 modifie plus de $50\%$ des acides aminés.

Dans la suite, lorsqu'une banque est modifiée à l'aide d'une matrice PAMx, nous concatènerons au nom de la banque la mention _PAMx. Ainsi, la banque 18mix modifiée à l'aide d'une matrice PAM20 sera nommée 18mix_PAM20.

## 5.3 Critères d'évaluation

Les logiciels d'interprétation des données de spectrométrie de masse doivent relever un double défi. En effet, les appareils disponibles aujourd'hui sont capables de générer des milliers de spectres MS/MS en quelques heures. Les logiciels développés doivent donc faire face à cet afflux de données. Par ailleurs, les spectres expérimentaux sont affectés par différentes sources de variabilité qui masquent le signal biologique. L'identification des peptides puis des protéines est donc délicat, et il faut être très attentif à la précision des résultats obtenus. Ce sont donc sur ces deux critères de temps d'exécution et de précision des résultats obtenus que nous avons décidé d'évaluer notre approche, les deux critères étant parfois antagonistes et la recherche du meilleur équilibre difficile à trouver.

### 5.3.1 Temps d'exécution

L'algorithme PacketSpectralAlignment est un algorithme lent au regard des approches développées dans des outils basés sur des méthodes de type SPC ne cherchant pas de modifications. Cette approche est compensée par des résultats précis qui vont permettre d'identifier des peptides modifiés. Cependant, il serait déraisonnable d'utiliser cette approche de manière systématique sur l'ensemble des spectres issus d'une analyse. L'utilisation de l'algorithme PacketSpectralAlignment doit s'intégrer dans des stratégies d'interprétation des données en plusieurs étapes consécutives associées à plusieurs outils logiciels, l'utilisation de notre méthode se limitant à l'interprétation des spectres de qualité correcte qui n'ont pas pu être identifiés par des approches plus rapides.

Différentes heuristiques peuvent cependant améliorer le temps de réponse sans dégrader de manière notable la précision des résultats. Pour chacune des heuristiques que nous avons testées, une mesure précise de son impact sur les **temps d'exécution** a été effectuée pour retenir les meilleurs compromis.

### 5.3.2 Qualité des résultats

Nous avons parlé plusieurs fois d'évaluer la **qualité** des résultats obtenus. Il existe différentes manières de le faire. Nous avons choisi d'évaluer la méthode d'identification comme un classifieur qui pour chaque peptide indique si l'identification est **attendue** ou non.

Définir un **résultat attendu** n'est cependant pas trivial. Tout d'abord, nous devons disposer d'un jeu de spectres ayant un peptide connu associé à chacun des spectres, comme par exemple le jeux de spectres spectres_ISB. Nous définirons ensuite une **identification attendue**, comme étant une comparaison attribuant au spectre le même peptide que celui qui était associé au spectre dans le jeu de données. Cependant, cela ne fonctionne que dans le cas où aucune modification n'est attendue. Dans le cas de modifications, nous devons calculer la distance séparant les peptides de la banque des peptides associés aux spectres du jeu de données (voir Section 5.2.1.1, page 62). Cette distance est calculée en utilisant l'algorithme d'alignement global de Needleman-Wunsch [NW70]. Si cette distance, qui caractérise le nombre de modifications séparant les deux peptides, est inférieure à une valeur $D$, alors nous considérons le peptide de la banque comme un résultat attendu. Cette valeur $D$ sera ajustée ultérieurement de sorte à

s'accorder avec le nombre de modifications recherchées par la méthode.

En opérant de la sorte, nous pouvons aisément utiliser une *Receiver Operating Characteristic* (ROC) *curve*, que nous appellerons **courbe ROC**. Une courbe ROC représente le tracé du taux de faux positifs en fonction du taux de vrais positifs. Le **taux de faux positifs** représente le nombre d'identifications présumées correctes alors qu'en réalité elles ne sont pas attendues ; tandis que le **taux de vrais positifs** correspond au nombre d'identifications présumées correctes qui sont effectivement attendues.

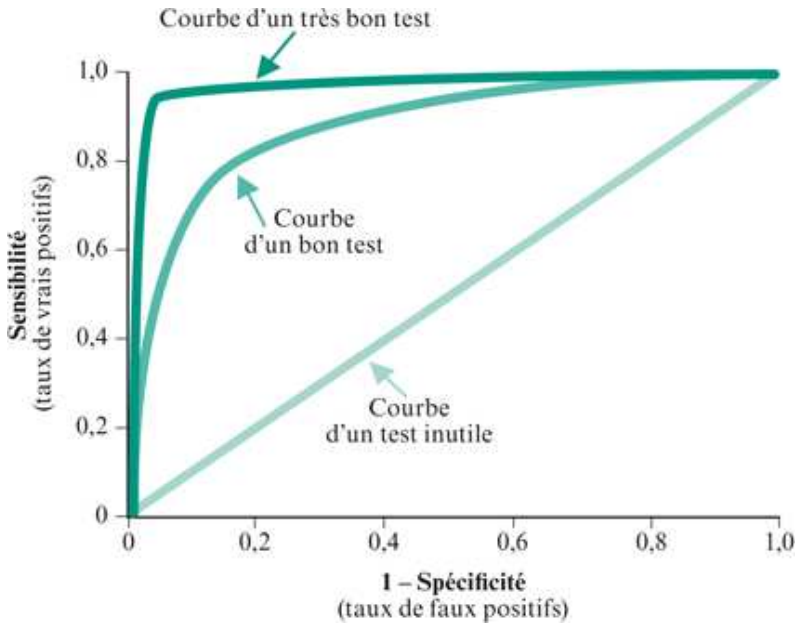


Figure 5.3 – Exemple de courbes ROC. *(Source :www.medhyg.ch)*

La Figure 5.3 présente 3 courbes ROC différentes sur le même tracé. Pour juger rapidement des résultats donnés par des courbes ROC, nous utilisons une seule mesure : l'aire sous la courbe (nommée **AUC**). L'AUC est une valeur comprise entre 0 et 1, plus elle tend vers 1, plus la méthode est de bonne qualité. Une courbe ROC qui suit la diagonale (AUC = 0,5) correspond à un classifieur aléatoire (même proportion de succès que d'échecs), tandis qu'une AUC inférieure à 0,5 signifiera que le classifieur classe à l'inverse de ce qui lui est demandé.

**Le cas d'une protéine attendue, mais sans association spectre-peptide.** Tous les jeux de données n'associent pas de manière sûre un peptide à chacun des spectres. Dans certains cas, seule la protéine que la méthode est censée retrouver est connue avec certitude. Dans un tel cas, nous considérerons comme mesure de qualité le nombre de peptides identifiés appartenant à la protéine attendue. Plus ce nombre de peptides est important, meilleure sera la qualité.

# CHAPITRE 6

# SIFpackets : mettre PacketSpectralAlignment en situation réelle

## 6.1 Introduction

Nous allons décrire, dans ce chapitre, les approches empiriques que nous avons développées pour ajuster les paramètres de l'algorithme PacketSpectralAlignment de manière à exploiter au mieux les potentiels de cet algorithme. Qu'il s'agisse de prétraitements, de filtrage de données ou de l'élaboration des scores d'ordonnancement des résultats, l'objectif est de concentrer les temps de calcul de notre algorithme sur l'information utile, tout en essayant de préserver, voire d'améliorer, la précision de notre méthode d'identification. L'intégration de tous ces éléments dans un même framework, appelé SIFpackets, permet d'aboutir à l'identification des protéines. Les résultats expérimentaux obtenus sur plusieurs jeux de données soulignent l'intérêt du framework que nous avons développé.

Une partie de ces travaux, réalisés en collaboration avec Guillaume Fertin, Irena Rusu et Dominique Tessier, ont été publiés dans [CFRT10].

## 6.2 Amélioration de l'identification des peptides : paramétrage et prétraitements

Les appareils de spectrométrie de masse sont capables aujourd'hui de générer des milliers de spectres MS/MS en quelques heures. Les algorithmes de traitement doivent être adaptés à une telle volumétrie. Pour améliorer la vitesse d'interprétation des spectres et ne garder que le signal utile, un paramétrage adapté de l'algorithme PSA, différents traitements des spectres et de la banque de donnée de référence peuvent être très efficaces. Dans cette section, nous détaillerons dans un premier temps les traitements que nous proposons sur les spectres expérimentaux, puis ceux sur les spectres théoriques. Nous décrirons ensuite le paramétrage de l'algorithme de comparaison de deux spectres.

### 6.2.1 Filtrage des spectres expérimentaux

Après les premiers prétraitements appliqués au signal sortant du spectromètre de masse, un spectre expérimental comporte un nombre de pics extrêmement variable, pouvant aller d'une dizaine à plusieurs centaines de pics. La vitesse d'exécution des algorithmes de comparaison

de spectres comme PacketSpectralAlignment est dépendante du nombre de pics traités lors de l'alignement. Pour que notre méthode PacketSPectralAlignment soit performante, il faut donc limiter le nombre de pics à aligner au strict nécessaire, en prenant toutefois garde de ne pas dégrader la précision des résultats, ce qui pourrait se produire en cas de retrait de pics utiles.

**Le nombre de pics à retenir est fonction du nombre d'acides aminés du peptide.** Il serait possible de limiter le nombre de pics d'un spectre à un nombre constant, ce qui reviendrait purement et simplement à supprimer les plus petits pics d'un spectre jusqu'à ce qu'il ne reste plus que le nombre de pics désirés. Cette méthode très simple est fréquemment employée, et même si cette solution ne donne pas de mauvais résultats pour des méthodes comme Mascot ou Sequest, comme ont pu le montrer Renard et al. dans [RKM+09], elle n'est certainement pas la plus adaptée. En effet, le nombre de pics escompté dans le spectre est dépendant du nombre d'acides aminés qui composent le peptide représenté. Il paraît donc logique de filtrer le spectre de sorte à ce qu'il conserve un nombre de pics proportionnel à sa longueur. En effet, conserver trop ou trop peu de pics dans un spectre peut s'avérer pénalisant pour les résultats de l'alignement.

**Un "beau" spectre est un spectre avec une répartition homogène des pics le long des fragmentations d'un peptide.** Un filtrage par fenêtre permet de limiter le nombre de pics dans une fenêtre glissant le long du spectre. La largeur d'une fenêtre correspond à un delta de masse donnée. Un filtrage par fenêtre permet d'ôter des pics dans les zones denses en pics -comportant vraisemblablement trop d'information par rapport à celle que l'on est capable d'utiliser dans l'algorithme d'alignement-, sans pour autant toucher aux zones pauvres en pics -ne comportant pas ou peu d'information superflue-. Un filtrage par fenêtre se caractérise par deux paramètres importants : la largeur de la fenêtre et le nombre de pics conservés dans l'intervalle défini par cette fenêtre. Une fenêtre de largeur égale à 110 daltons est présentée en bleu dans la Figure 6.1.

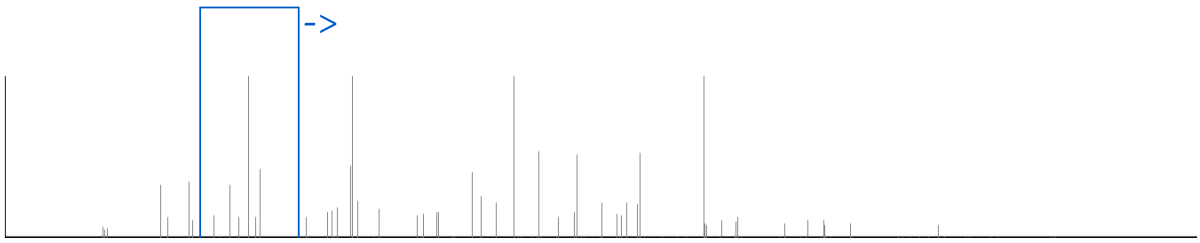


Figure 6.1 – Un spectre MS/MS bruité sur la gauche et avec peu de pics sur la droite. Le rectangle bleu représente une fenêtre de filtrage de largeur égale à 110 daltons.

La méthode de filtrage avec fenêtre présente l'avantage de réduire le nombre de pics de manière proportionnelle à la longueur du spectre, via un voisinage local dans le spectre. Cela est particulièrement intéressant pour éliminer le bruit, qui est souvent localisé dans une zone particulière du spectre comme illustré dans l'exemple de la Figure 6.1, où un filtrage par fenêtre

est particulièrement bien adapté.

Les travaux de Renard et al. [RKM+09] ont révélé que pour des méthodes comme Mascot ou Sequest, la méthode de filtrage par fenêtre donne d'excellents résultats. Cependant, il faut être prudent quant aux paramètres à utiliser pour ce filtrage, ceux-ci pouvant dépendre de la méthode de comparaison qui est utilisée par la suite.

L'objectif du filtrage par fenêtre est de faire ressortir les pics les plus importants marquant un site de fragmentation. Ainsi, une fenêtre de taille idéale devrait toujours contenir un seul site de fragmentation. Cependant, étant donné la grande disparité dans la taille des acides aminés (cf. Tableau 2.2, page 10) il n'est pas possible de définir une taille de fenêtre permettant cela. Un compromis raisonnable consiste à utiliser la taille moyenne d'un acide aminé en tant que largeur de fenêtre. Cette taille moyenne, en tenant compte de l'abondance relative des différents acides aminés, est de 111,27 daltons. Nous avons donc choisi une fenêtre de largeur égale à 110 daltons.

La Figure 6.2 permet d'évaluer l'impact de la taille de la fenêtre utilisée sur la qualité des résultats ainsi que sur le temps d'exécution. Le graphique de gauche nous apprend qu'une fenêtre de petite taille peut être néfaste pour la qualité des résultats, dans les cas où 6 et 8 pics sont conservés, tandis que pour une fenêtre comprise entre 80 et 120 daltons, la qualité ne varie que très peu. La qualité se verra de nouveau dégradée pour une fenêtre de plus de 120 daltons. Le graphique de droite nous apprend quant à lui que le temps d'exécution est directement proportionnel à la taille de la fenêtre. Fixer la largeur de fenêtre à 110 apparaît donc ici comme un bon compromis, et ce quelque soit le nombre de pics conservés, que ce soit en termes de qualité ou de temps d'exécution.

Pour ce qui est du nombre de pics à conserver, nous avons dû tester différentes possibilités, car en utilisant la notion de paquet pour l'alignement, le nombre de pics à conserver n'est peut être pas le même que pour des méthodes comme Mascot ou Sequest. Il est à noter que le filtrage du bruit est effectué sur le spectre expérimental brut, avant de le transformer en spectre expérimental symétrique.

La Figure 6.3 présente une évaluation du filtrage par fenêtre, pour différentes quantités de pics conservés dans une fenêtre de 110 daltons, sur le jeu de données spectres_ISB et la banque 18mix_PAM40. La courbe noire évalue la qualité via l'AUC des courbes ROC, et la courbe grise évalue le temps d'exécution de la méthode. Nous pouvons noter que :

- Le filtrage a un impact direct sur la qualité. La conservation de trop, ou de trop peu de pics réduit sensiblement la qualité des résultats. Il apparaît opportun de conserver entre 6 et 7 pics par fenêtre de 110 daltons.
- Le temps d'exécution est directement proportionnel au nombre de pics conservés, il est donc utile d'éliminer le maximum de pics superflus.

Nous avons décidé de **conserver six pics dans une fenêtre de 110 daltons**, car même si la qualité nous paraît plus importante que le temps d'exécution, un nombre de pics supérieur à six améliore trop peu la qualité, alors que la nuisance en temps d'exécution se fait fortement ressentir.

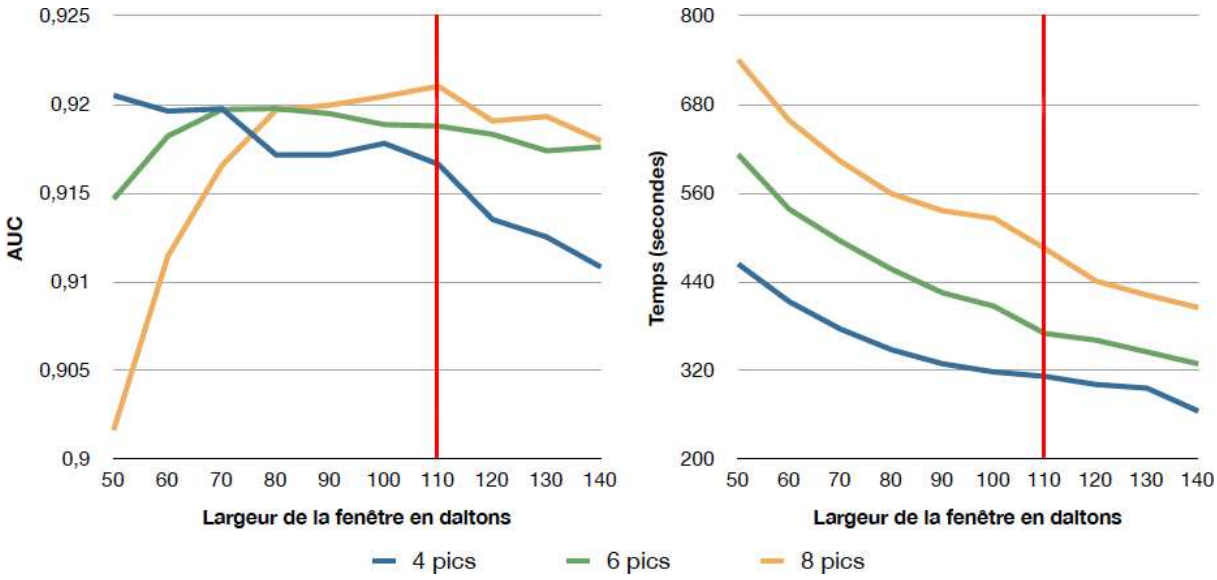


Figure 6.2 – Évaluation du comportement de l'algorithme PacketSpectralAlignment -AUC et temps d'exécution- en fonction de la largeur de la fenêtre et du nombre de pics conservés dans la fenêtre -4, 6 ou 8 pics- avec le jeu de données spectres_ISB et la banque 18mix_PAM40.

### 6.2.2 Modèle de score pour l'alignement d'un paquet

Pour mesurer la ressemblance entre deux spectres $SS_t$ et $SS_e$, il est nécessaire d'attribuer un score qui tienne compte de la notion de paquet ainsi que de la symétrie. Le score global de comparaison de deux spectres peut correspondre à l'addition des scores individuels obtenus par l'alignement des paquets sur les positions possibles. Ce score est utilisé par la fonction `score`(Paquet $p_i$, Position Possible $pp_j$) dans l'Algorithme 1, page 57, où $p_i$ est un paquet de $SS_t$ et $pp_j$ une position possible de $SS_e$.

Il existe de très nombreuses possibilités pour calculer un score de ce type, la plus simple étant d'appliquer un Shared Peaks Count (SPC), c'est-à-dire de compter le nombre de pics en commun entre le paquet $p_i$ de $SS_t$ et le spectre expérimental symétrique $SS_e$ quand le paquet $p_i$ est aligné sur la position possible $pp_j$. Ce score trivial peut être amélioré en prenant en compte la structure même des paquets. En effet, dans un paquet nous savons quel type d'ion est représenté par chacun des pics, et cette information peut être couplée aux probabilités d'apparition des ions dans un spectre. Ces probabilités ont été obtenues de manière empirique par différentes équipes et publiées dans [DAC+99, HHS03, FP05]. Le Tableau 6.1 met en évidence l'intérêt de pondérer le score SPC en utilisant pour chacun des pics des paquets, leur probabilité d'apparition (le détail de la pondération est donnée dans le Tableau 6.2). Il est à noter que cette pondération n'influe pas sur le temps d'exécution.

Il est fréquemment observé, dans les méthodes de scores, que prendre en compte l'intensité des pics améliore considérablement la qualité des résultats. Cela permet en effet de mieux attribuer les ions $b$ et $y$ qui sont généralement représentés par des pics de forte intensité lors

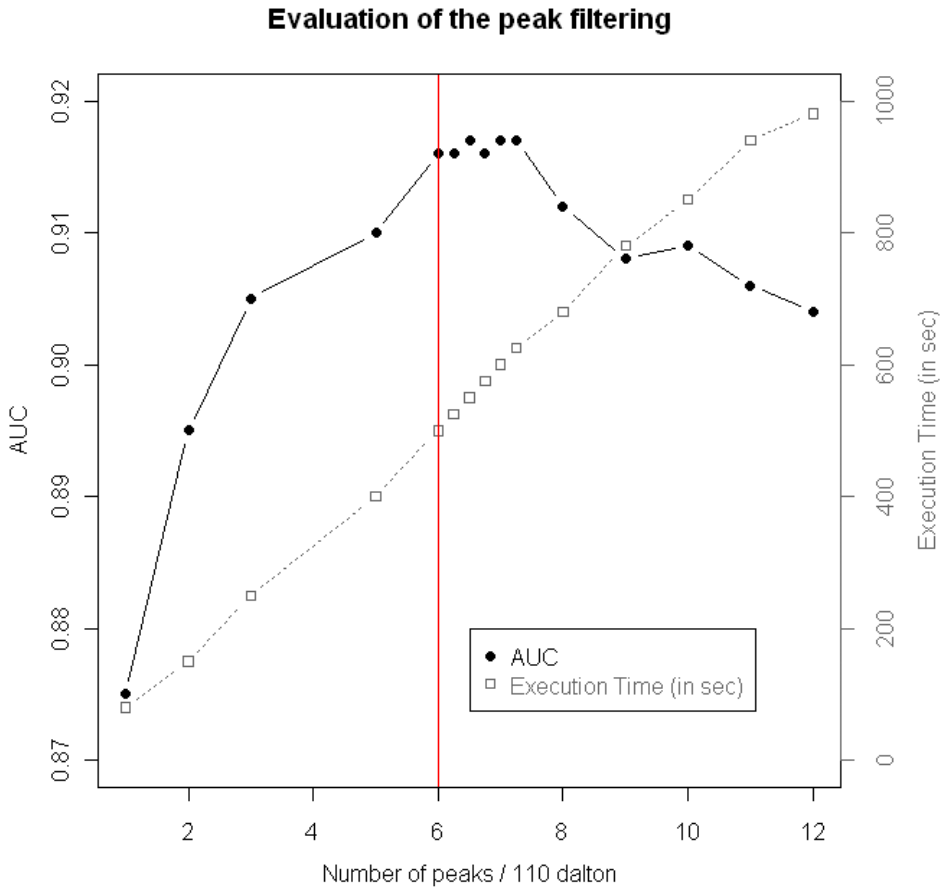


Figure 6.3 – Évaluation de l'efficacité d'un filtrage par fenêtre en fonction du nombre de pics conservés dans une fenêtre constante de largeur 110 daltons, sur le jeu de données spectres_ISB et la banque 18mix_PAM40.

de l'alignement. Nous avons évalué différentes méthodes afin d'ajouter l'intensité des pics dans notre modèle de score. Lors de ces expérimentations, le nombre de pics dans un spectre expérimental était restreint aux 100 pics les plus intenses et la probabilité d'apparition des ions était utilisée en plus de la prise en compte de l'intensité. Les scores S1 à S7 ont été élaboré de la manière suivante :

- Le score S1 ne tient pas compte de l'intensité.
- Pour calculer les scores S2 à S5, les pics du spectre sont ordonnés par intensité décroissante. À chaque pic correspond son rang de classement, le rang 0 étant attribué au pic le plus intense. Les scores S2 à S5 utilisent une constante (entre 100 et 400 selon les cas) de laquelle est ôté le rang du pic.
- Le score S6 correspond à une normalisation de l'intensité, où le plus grand pic a une intensité de 1 et le plus petit une intensité de 0.
- Enfin, le score S7 est aussi dépendant du rang des pics, et donne une intensité comprise

| **Méthode de score** | **AUC** |
|---|---|
| SPC | 0.915 |
| SPC pondéré | 0.920 |

Table 6.1 – Comparaison de l'AUC de la méthode PacketSpectralAlignment selon l'utilisation d'un score de type SPC ou d'un score prenant en considération les probabilités d'apparition des différents ions (SPC pondéré), sur le jeu de données spectres_ISB et la banque 18mix_PAM40.

| **Nom du fragment** | **Probabilité** | **Pondération de l'ion** |
|---|---|---|
| $y$ | 0,87 | 8,7 |
| $y^*$ | 0,24 | 2,4 |
| $y^\circ$ | 0,26 | 2,6 |
| $b$ | 0,83 | 8,3 |
| $b^*$ | 0,36 | 3,6 |
| $b^\circ$ | 0,39 | 3,9 |
| $a$ | 0,34 | 3,4 |
| $a^*$ | 0,20 | 2,0 |
| $a^\circ$ | 0,17 | 1,7 |

Table 6.2 – Liste des fragments du modèle paquet avec pour chacun d'eux la pondération utilisée dans le score.

entre 1 et 2, mais qui est en pratique très proche de 1 pour la majorité des pics. Ce système permet juste d'accorder plus de poids aux quelques pics les plus intenses, ces pics traduisant une information très importante.

Les différents résultats obtenus sont présentés dans la Figure 6.4, avec le jeu de données spectres_ISB et la banque 18mix_PAM40. De manière empirique, nous pouvons observer que la prise en compte de l'intensité dans nos différents essais dégrade la valeur de l'AUC. Nous pouvons expliquer ce comportement grâce à la notion de paquet. En effet, si la notion de paquet remplit bien son rôle, le positionnement du paquet est optimal et en conséquence, il englobe de nombreux pics d'une même fragmentation, dont ceux de forte intensité. En effet, les pics de plus forte intensité sont généralement des ions $b$ ou des ions $y$ (comme ont pu le montrer Dancik et al. dans [DAC+99]), des pics faisant partie de notre modèle de paquet. La prise en compte de l'intensité ne changerait donc rien à l'alignement, les pics de forte intensité étant déjà alignés. La dégradation des résultats peut aussi s'expliquer. La prise en compte de l'intensité des pics peut induire une forte différence dans le poids accordé aux différents pics d'un même paquet. Il en ressort donc que le pic le plus intense va contribuer de manière majoritaire au score attribué au paquet, limitant les autres pics à une participation marginale. De plus, un petit pic isolé n'aura pas de poids dans le score global, alors que le fait qu'il soit isolé peut justement porter beaucoup d'information. Dans ce dernier cas, l'intensité devrait être évaluée localement dans le spectre, et non globalement.

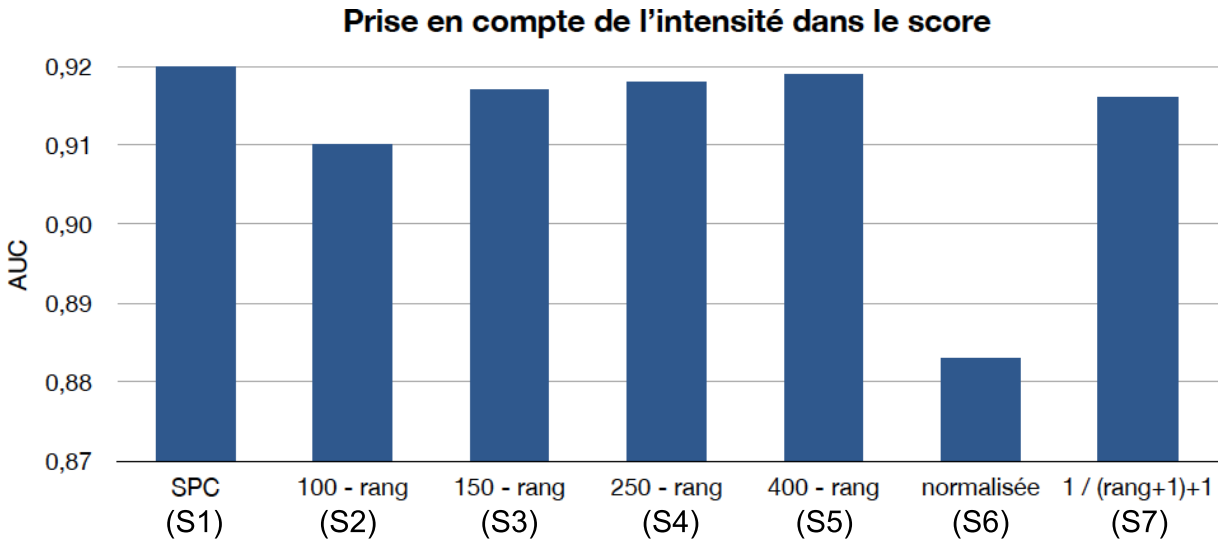


Figure 6.4 – Évaluation de différentes méthodes de score, prenant en compte l'intensité des pics, pour l'alignement d'un paquet en utilisant le jeu de données spectres_ISB et la banque 18mix_PAM40.

Au vu de ces différents résultats, nous avons choisi d'utiliser un score basé sur une pondération des différents types d'ions à l'aide de leur probabilité d'apparition dans le spectre, et de ne pas tenir compte de l'intensité des pics, l'apport étant inintéressant du fait de l'usage des paquets. La Figure 6.5 donne différents exemples de scores résultant de l'alignement d'un paquet.

### 6.2.3 Filtrage des positions possibles

Dans la Section 4.4.3.2, page 53, nous avons introduit la notion de position possible qui définit, en fonction de l'alignement, l'ensemble de toutes les positions sur lesquelles un paquet peut s'aligner. Plus les contraintes d'alignement sont importantes et plus l'ensemble des positions possibles est réduit, ce qui améliore le temps d'exécution. En revanche, si les contraintes d'alignement sont trop fortes, l'algorithme perd des capacités d'alignement et risque de perdre en qualité des résultats.

Tout d'abord, nous utilisons l'Algorithme 3 pour établir la liste des positions possibles. Plusieurs points importants sont à noter dans celui-ci :

- Nous parcourons toutes les positions $m$ de $SS_e$, ce qui signifie toutes les positions envisageables selon le niveau de précision souhaité par la méthode.
- À la ligne 3 de l'Algorithme 3, un paquet $p$ est aligné sur la position courante $m$ pour évaluer l'apport de la position $m$ si elle venait à faire partie d'un alignement. La fonction `score()` utilise le score tel que nous l'avons défini précédemment dans la Section 6.2.2.
- Enfin, à la ligne 5 de l'Algorithme 3, la position courante $m$ est conservée uniquement si son apport dans un alignement, donné par $score$, est supérieur à un seuil $SEUIL$ donné. Cette variable $SEUIL$ présente deux intérêts. Tout d'abord, elle permet de régler plus

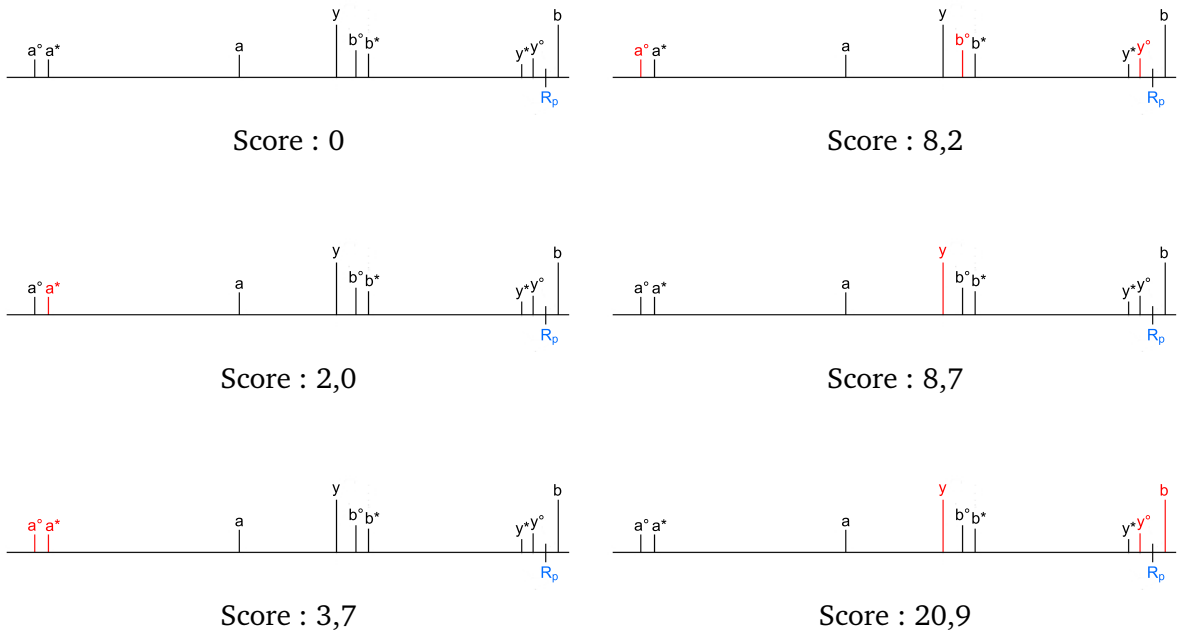


Figure 6.5 – Différents exemples de scores pour différents alignements de paquets. Ne sont représentés ici que les paquets, avec en rouge les pics alignés et en noir les pics non alignés.

finement le choix des positions possibles à conserver ou non, et donc d'influer directement sur la qualité et le temps d'exécution, mais surtout, elle permet un réglage rapide dans le cas d'un changement de méthode de score. En effet, le seuil utilisé dépend uniquement de la méthode de score utilisée par la fonction `score()`.

Nous avons testé différents seuils de sélection des positions possibles, sur le jeu de données spectres_ISB et la banque 18mix_PAM40, en modifiant notre variable $SEUIL$ pour des valeurs appartenant à l'intervalle $[0; 20]$. Ces valeurs ont été choisies pour notre intervalle en fonction du score que nous avons défini :

- la variable $SEUIL$ à $0$ conserve toutes les positions possibles qui contribuent au score, c'est-à-dire toutes les positions pour lesquelles au moins un pic s'aligne avec un paquet.
- la variable $SEUIL$ à $20$ conserve toutes les positions possibles ayant au moins un ion $x$, un ion $y$ et d'autres ions moins importants alignés avec un paquet. Cela signifie donc une très forte contrainte sur l'alignement de la position en question.

La Figure 6.6 permet de visualiser la qualité du filtrage en fonction de la variable $SEUIL$ en terme d'AUC (sur la courbe noire) ainsi qu'en terme de temps d'exécution (avec la courbe grise) sur le jeu de données spectres_ISB et la banque 18mix_PAM40. Nous pouvons donc remarquer que sur l'intervalle $[0, 8]$, la qualité des résultats mesurée par l'AUC reste stable :

- Cela peut s'expliquer par le fait qu'un seuil de 8 est dans notre modèle de score la contribution d'un ion $b$ ou d'un ion $y$, or ces ions sont censés être présents avec une très forte probabilité. En revanche, augmenter la valeur de $SEUIL$ au-delà de ce niveau va réduire la qualité des résultats. En effet, cela signifie qu'outre un ion $b$ ou $y$, nous demandons

**Algorithm 3** Algorithme de création et de filtrage de la liste des positions possibles

**Entrée :**
Spectre expérimental $SS_e$ et
Réel $SEUIL$

**Sortie :**
Ensemble de Positions Possibles de $SS_e$

1: $PP = \emptyset$
2: **pour tout** position $m$ de $SS_e$ **faire**
3: $s =$ `score`(Packet $p$, $m$)
4: **si** $s \geq SEUIL$ **alors**
5: $PP = PP \cup m$
6: **fin si**
7: **fin pour**
8: **return** $PP$

obligatoirement d'autres pics pour attester de l'intérêt d'une position, or leur présence ne se vérifie pas systématiquement dans le spectre. La Figure 6.5 donnant certains exemples de scores pour différents alignements de paquets, illustre ce choix de variable $SEUIL$.
– D'un point de vue temps d'exécution, nous pouvons voir qu'il diminue très rapidement avec l'accroissement de la variable $SEUIL$ jusqu'à un $SEUIL$ de $10$ environ, puis ne diminue plus que très lentement par la suite. La raison en est que plus le seuil est élevé, moins il y a de positions dont le score dépasse cette valeur. Or le temps d'exécution est vraisemblablement directement proportionnel au nombre de positions possibles traitées par l'algorithme d'alignement.

En conséquence de ces résultats, **nous avons choisi de fixer la variable $SEUIL$ à $8$** pour garantir la meilleure qualité possible, tout en réduisant considérablement le temps d'exécution dans la suite de notre étude. Mais ce choix est discutable : par exemple, dans certains cas, fixer la variable $SEUIL$ à $10$ pourrait permettre un gain de temps non négligeable, pour une diminution de la qualité qui reste raisonnable. Le choix dépend surtout du contexte d'utilisation, ainsi que de la taille des données analysées, et peut donc être paramétré en fonction des besoins.

Dans les spectres expérimentaux, nous pouvons observer des zones beaucoup plus denses en positions possibles que d'autres. Il apparaît donc intéressant, après la première étape de sélection, de les filtrer à la manière des pics, à l'aide d'un filtrage par fenêtre. Le but est de conserver localement (c'est-à-dire dans un espace délimité par une fenêtre) un nombre limité de positions possibles constitué de celles offrant les contributions les plus importantes. Comme pour le filtrage par fenêtre des pics, nous devons définir deux paramètres : la taille de la fenêtre et le nombre de positions possibles à conserver pour chacune des positions de la fenêtre.

Une fenêtre de largeur idéale serait une fenêtre contenant systématiquement un unique site de fragmentation. Le choix de largeur est donc le même que celui effectué pour la fenêtre de

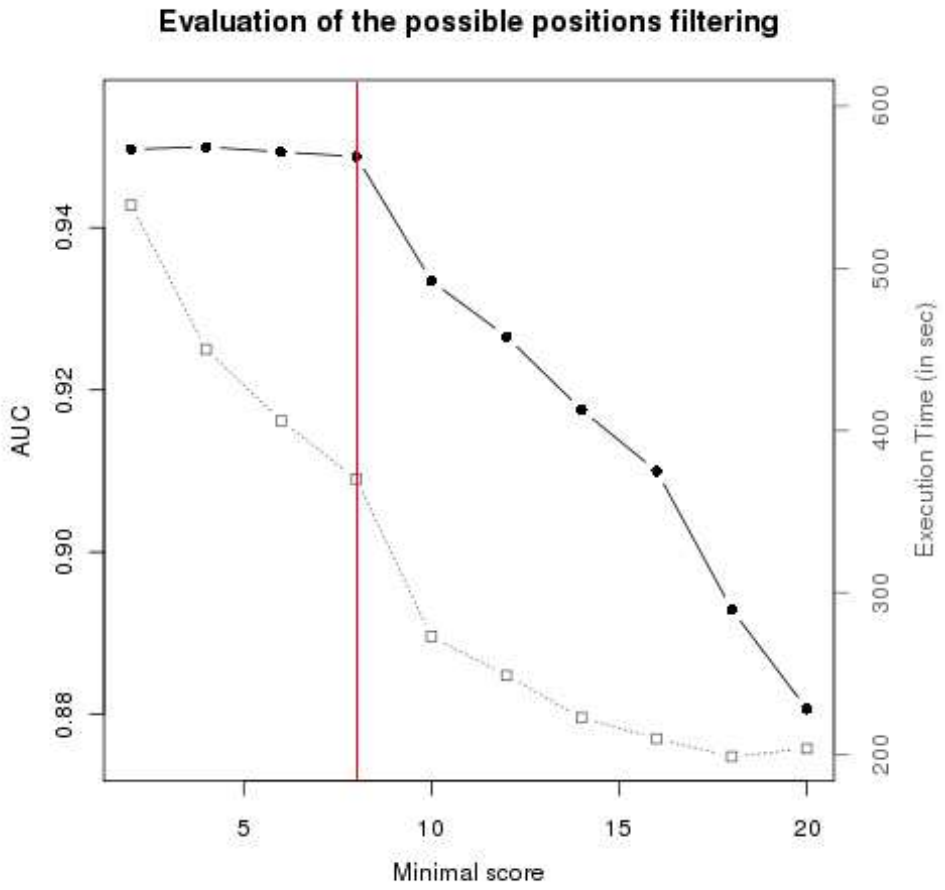


Figure 6.6 – Évaluation de l'impact du filtrage des positions possibles en terme de temps (courbe grise) et en terme de qualité (courbe noire) sur le jeu de données spectres_ISB et la banque 18mix_PAM40.

filtrage des pics du spectre expérimental (Section 6.2.1, page 69) : la masse moyenne d'un acide aminé, à savoir 110 daltons.

En revanche, des tests sont nécessaires pour définir le nombre de positions possibles à conserver dans l'intervalle défini par notre fenêtre. Car, même si en pratique l'algorithme n'est censé utiliser qu'une ou deux positions possibles dans cet intervalle durant l'alignement (deux positions dans le cas limite où nous sommes en présence de deux glycines, l'acide aminé le plus léger), rien ne garantit que ce seront nécessairement celles ayant la plus forte contribution qui devront être utilisées. Nous avons choisi de faire varier le nombre de positions possibles à conserver de 2 à 10, et d'en évaluer l'impact sur les résultats. Cette évaluation a été conduite sur le jeu de données spectres_ISB et la banque 18mix_PAM40. Nous voyons dans la Figure 6.7 qu'en terme de qualité (courbe noire), conserver trop peu de positions possibles est très néfaste à la méthode, mais également qu'en conserver une trop grande quantité peut, dans une moindre

mesure, en réduire la qualité. La courbe représentant l'impact sur le temps d'exécution rappelle bien que le temps de comparaison est directement proportionnel au nombre de positions possibles à traiter, et que donc en conserver le moins possible est intéressant. Nous avons ici choisi de fixer à 6 le nombre de positions possibles à conserver par fenêtre de 110 daltons, afin de maximiser la qualité du résultat.

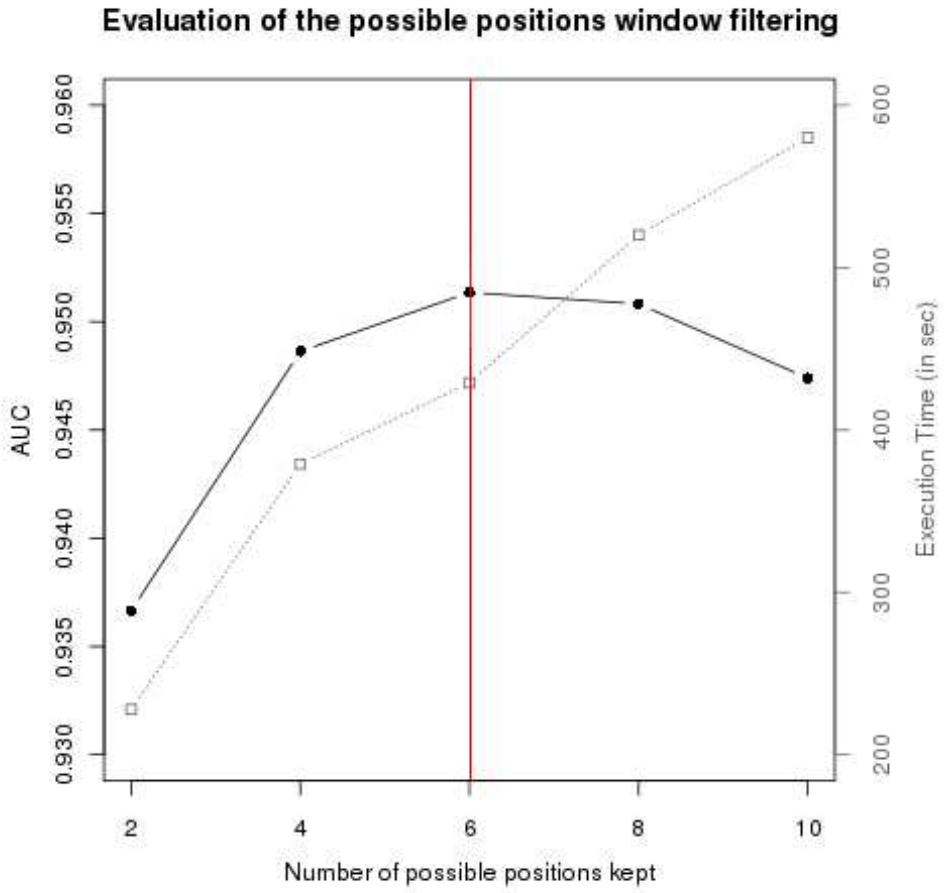


Figure 6.7 – Évaluation du filtrage des positions possibles à l'aide d'une fenêtre en terme de temps (courbe grise) et en terme de qualité (courbe noire) sur le jeu de données spectres_ISB et la banque 18mix_PAM40.

Nous avons pu démontrer ici que la sélection et le filtrage des positions possibles permet un gain modéré en terme de qualité, mais surtout une réduction très importante du temps d'exécution lors de la comparaison. Nous pouvons souligner que ce traitement, même s'il peut prendre un temps non négligeable, n'est à effectuer qu'une unique fois sur chacun des spectres, et qu'il fera gagner un temps très important lors de la comparaison.

### 6.2.4 Filtrage de la banque

Dans une approche par comparaison de spectres, la taille de la banque est certainement le facteur qui a le plus d'influence sur le temps d'exécution final. Lorsque le but recherché est une identification sans modification, la masse du peptide à rechercher dans la banque est connue. Il est donc possible de ne considérer, lors de la comparaison, que les peptides de masse similaire, ce qui réduit énormément le nombre de comparaisons et facilite l'identification. Cependant, dans le cas où des modifications sont présentes et recherchées, la masse mesurée du peptide analysé et la masse théorique d'un peptide "proche" dans la banque sont potentiellement différentes. Il n'est donc pas possible de restreindre les comparaisons aux peptides de masses similaires. En revanche, nous pouvons nous interroger sur le moyen de filtrer les peptides de cette banque afin de ne pas comparer un spectre avec des peptides trop éloignés du candidat attendu.

Pour répondre à ce problème, nous proposons de filtrer les peptides de la banque de telle sorte que leur masse soit proche de la masse du peptide analysé. Nous avons défini cette distance **proportionnellement au nombre d'acides aminés d'un peptide**, car plus un peptide contient d'acides aminés, plus il a de chance de porter des modifications.

Nous pouvons considérer qu'une modification dans un peptide ne va pas faire varier la masse de ce peptide d'un facteur trop important, et ce pour plusieurs raisons :

- Dans le cas d'une substitution d'acides aminés, la différence de masse observée sera la différence de masse entre l'acide aminé substitué et l'acide aminé substituant. Dans le cas le plus extrême, la substitution aura lieu entre une glycine (de masse 57 Da) et un tryptophane (de masse 186 Da) soit 129 Da d'écart quand la masse moyenne d'un acide aminé est de 110 Da. Cependant, d'autres considérations sont à prendre en compte. Tout d'abord il est peu probable qu'une glycine soit substituée par un tryptophane (la probabilité est de 0,0055 d'après [VST03]). En effet les acides aminés se substituent généralement par des acides aminés aux caractéristiques physico-chimiques proches, ce qui se traduit généralement par des substitutions d'acides aminés n'ayant pas des masses trop éloignées les unes des autres. De plus, dans le cas où il y a plusieurs substitutions, il est très probable qu'elles se compensent partiellement, c'est-à-dire qu'une substitution peut augmenter la masse du peptide, tandis qu'une seconde va la réduire, donnant un changement de masse assez faible pour le peptide.
- Dans le cas des modifications post-traductionnelles, la situation est comparable, car si en théorie elles peuvent faire varier la masse d'un peptide de près de 1000 Da, en pratique le changement effectué dépasse rarement la centaine de daltons [$M^+98$].
- L'insertion et la suppression d'acides aminés est un phénomène plus difficile à gérer. Le changement de masse engendré peut être extrêmement important, dans un sens comme dans l'autre. La difficulté étant de connaître la distance minimum à conserver entre les deux séquences comparées pour considérer qu'elles sont issues d'une séquence commune.

La Figure 6.8 permet d'évaluer l'impact d'un filtrage de la banque sur la qualité de notre méthode de comparaison sur le jeu de données spectres_ISB et la banque 18MIC_PAM40. Tout d'abord, cela fait apparaître que filtrer la banque améliore énormément la qualité des résultats, ce qui peut s'expliquer par une forte diminution des identifications aléatoires. Ensuite, cela montre que tolérer au plus $20\%$ d'écart de masse entre les deux spectres comparés est le plus

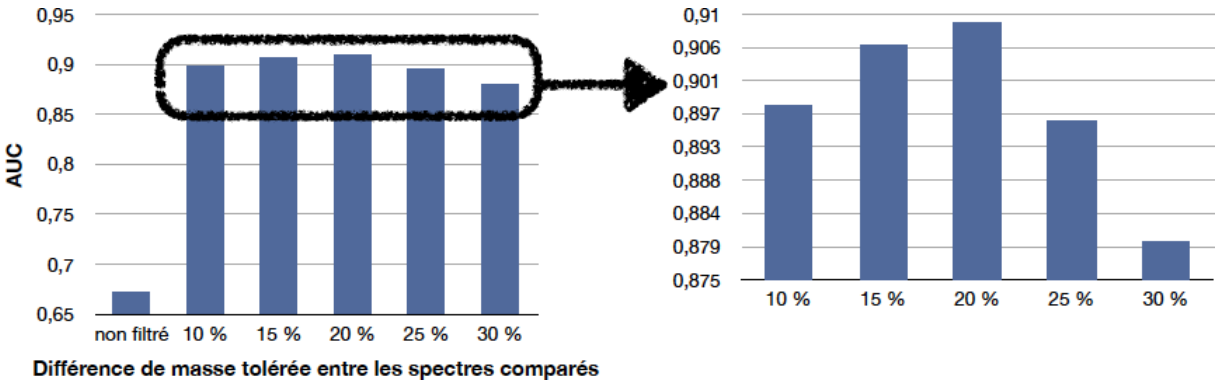


Figure 6.8 – Évaluation de la qualité de la méthode de comparaison pour un filtrage de la banque autorisant au maximum $10\%$ à $30\%$ de différence de masse entre le spectre expérimental et les candidats de la banque. L'AUC lorsque aucun filtrage n'est effectué est aussi donnée à titre indicatif. Cette évaluation a été effectuée sur le jeu de données spectres_ISB et la banque 18MIC_PAM40.

adapté. Nous avons donc choisi de conserver cette valeur de $20\%$ pour filtrer la banque dans la suite de nos travaux.

### 6.2.5 Nombre de modifications tolérées

L'algorithme PacketSpectralAlignment tolère un nombre maximum de modifications par peptide. Ce nombre, que nous nommons $K$, doit être défini par l'utilisateur au moment de l'utilisation de la méthode.

Le paramètre $K$ a un impact direct sur les résultats, qui se traduit par :

- Une forte influence sur la qualité des résultats. Un $K$ trop petit, en restreignant fortement le nombre de modifications tolérées, risque d'empêcher de nombreuses identifications. Cependant, un $K$ trop grand, en autorisant un nombre très élevé de modifications, risque de produire un nombre d'identifications extrêmement important, qui ne seront pas nécessairement pertinentes pour l'utilisateur.
- Une variation du temps d'exécution de la méthode, car l'algorithme PacketSpectralAlignment recherche nécessairement jusqu'à $K$ modifications. Ainsi le temps d'exécution va croître avec le paramètre $K$.

Un utilisateur à la recherche d'un faible nombre de modifications, par exemple à la recherche de modifications post-traductionnelles, peut définir $K$ comme une valeur constante assez petite (par exemple 1 ou 2), de manière comparable à ce qui est fait avec SpectralAlignment [PDT00, PMDT01].

En revanche, si l'utilisateur est à la recherche d'un nombre plus important de modifications, par exemple lors de la comparaison d'orthologues, il est intéressant de pouvoir utiliser de plus grandes valeurs de $K$. Pour ce faire, il est intéressant de ne pas définir ce paramètre comme constant, mais comme proportionnel à la longueur des peptides. En effet, plus un peptide com-

porte d'acides aminés, plus il a de chance de comporter des mutations. De plus, il n'est pas raisonnable de tolérer trop de modifications sur les peptides de petite taille (accepter 4 modifications dans un peptide de 6 acides aminés paraît démesuré tandis que 4 modifications pour un peptide de 25 acides aminés paraît plus approprié).

Nous avons choisi, pour nos tests, de définir $K$ pour traiter des organismes non séquencés. Nous nous sommes donc appuyés sur une valeur de $K$ proportionnelle à la taille des spectres comparés. La Figure 6.9 présente les résultats d'une évaluation de la qualité des résultats de PacketSpectralAlignment en fonction de différents paramètres $K$, tous proportionnels à la taille des spectres. Un paramètre $K$ de $0,05$ signifie que l'on tolère $0,05$ modification par $100$ daltons. Ce comportement est évalué sur la banque 18mix_PAM40. Le paramètre $K$ est donc optimisé pour identifier des séquences orthologues ayant environ $70\%$ d'identité.

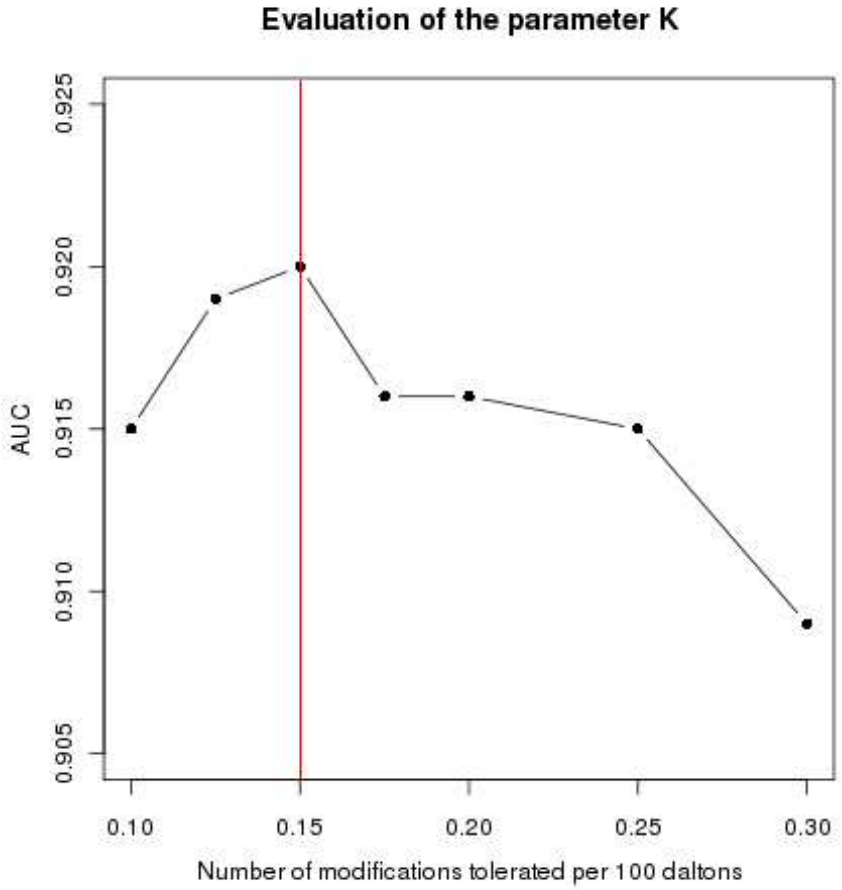


Figure 6.9 – Comparaison de différentes valeurs de modifications tolérées pour rendre $K$ dépendant de la longueur du peptide, avec le jeu de données spectres_ISB et la banque 18mix_PAM40.

La Figure 6.9 représente la qualité des résultats obtenus en fonction du paramètre $K$ choisi.

Nous observons ici qu'autoriser $0.15$ modification par $100$ daltons, c'est-à-dire 1 modification par 660 daltons ou encore 1 modification pour 6 acides aminés en moyenne, permet d'obtenir les meilleurs résultats. Avec moins de modifications tolérées, des alignements sont manqués, tandis qu'avec plus de modifications tolérées nous observons trop d'alignements non désirés.

### 6.2.6 Pénalités de modification

Le score d'un alignement que nous proposons est composé de l'addition des scores résultant de l'alignement de chacun des paquets du spectre théorique symétrique sur le spectre expérimental symétrique. Cependant, d'autres éléments sont à prendre en compte dans le score final de l'alignement. En effet, comme nous avons pu le définir dans l'algorithme PacketSpectralAlignment, nous appliquons des pénalités lorsque des modifications sont nécessaires pour obtenir un alignement. Il a donc fallu définir comment cette pénalité intervenait dans l'élaboration du score.

Cette pénalité vise à réduire le nombre de modifications appliquées au nombre minimal nécessaire, afin d'obtenir une identification correcte. Sans cette pénalité, notre méthode pourrait aisément introduire des modifications dans les zones fortement bruitées ou dans les zones présentant très peu de pics, dans le but d'améliorer le score même de manière très minime. Ces modifications sont généralement erronées.

La pénalité intervient au moment de l'alignement d'un paquet avec une des positions possibles, dans le cas où une modification est appliquée. Cette pénalité va réduire la contribution apportée au score par la position possible utilisée. Cela va donc permettre à la méthode de filtrer plus fortement les positions possibles utilisées dans le cas des modifications, et de réduire le score afin de privilégier un alignement sans modification à un alignement avec modification lorsque la différence en terme de score entre les deux est minime.

Le choix de cette pénalité dépend de différents éléments, tels que :

- le score utilisé et ce qui influence le score obtenu par un alignement, comme par exemple les caractéristiques de l'échantillon et du matériel,
- l'objectif de la recherche, car rechercher des modifications post-traductionnelles ou comparer des organismes orthologues est différent. Par exemple, lorsque moins de modifications sont attendues, il peut être intéressant d'augmenter la valeur de la pénalité.

Nous avons défini notre pénalité comme valant 10. Cela permet de ne privilégier l'alignement d'un paquet modifié sur un paquet non modifié que dans le cas où, après modification, le score obtenu est supérieur d'au moins 10. Nous avons choisi cette valeur car cela correspond à la présence d'un ion $b$ ou $y$ supplémentaire lors de l'alignement, ce qui renforce le choix de la modification par rapport aux alignements non modifiés potentiels.

La pénalité de modification n'a qu'un impact minime sur la qualité de l'identification d'un point de vue AUC. Elle permet surtout une meilleure discrimination parmi les peptides candidats ayant des scores élevés. De plus, nous avons pu constater que l'usage de la pénalité permet de mieux localiser les modifications lors de l'alignement, les résultats fournis après backtrack étant plus proches de ceux que l'on peut interpréter manuellement dans les spectres.

## 6.3 SIFpackets : une plate-forme complète associant spectres et peptides

### 6.3.1 Description de la plate-forme SIFpackets

À partir des différents éléments de filtrage permettant d'améliorer le fonctionnement de l'algorithme PacketSpectralAlignment, nous avons créé une plate-forme d'identification de spectres que nous avons nommé **SIFpackets**.

Cette plate-forme, décrite dans la Figure 6.10, combine tous les éléments vus précédemment afin de fournir les meilleurs résultats en terme de qualité comme de temps d'exécution, lors de la comparaison des spectres expérimentaux avec les peptides issus d'une banque de protéines. Comme un spectromètre de masse ne peut mesurer la masse des peptides que dans un intervalle de masses donné, SIFpackets filtre (Figure 6.10 (b)) les peptides issus de la banque en fonction de cette tolérance. Comme nous l'avons vu dans la Section 6.2.1, page 69, les spectres expérimentaux (Figure 6.10 (f)) sont aussi filtrés (Figure 6.10 (g)) afin de limiter, entre autres, le nombre de pics qu'ils comportent.

Ensuite, les deux spectres sont modifiés pour prendre en compte la symétrie (Figure 6.10 (d) et (h)) comme cela a été expliqué dans la Section 4.4, page 48. Puis les positions possibles sont créées et filtrées (Figure 6.10 (i)) comme expliqué dans la Section 4.4.3.2, page 53, pour la création et dans la Section 6.2.3, page 75, pour le filtrage.

Il est alors possible de comparer les spectres expérimentaux symétriques et les spectres théoriques symétriques à l'aide de l'algorithme PacketSpectralAlignment (Figure 6.10 (k)). PacketSpectralAlignment permet de faire une identification rapide des peptides en présence de modifications.

Même si la plate-forme SIFpackets est construite autour de la méthode PacketSpectralAlignment, elle n'en est pas moins adaptable et utilisable pour améliorer les résultats d'une autre méthode de comparaison de spectres. Certaines étapes présentées ici, notamment le filtrage de la banque ou des spectres expérimentaux, peuvent être appliquées à n'importe quel autre algorithme de comparaison.

### 6.3.2 Variante permettant une meilleure prise en compte des modifications

La méthode d'alignement PacketSpectralAlignment décrite dans le Chapitre 4 fonctionne comme escompté, cependant, elle ne tient pas compte d'un élément. Nous avons en effet pu constater qu'avec l'application de la symétrie sur tous les pics dans le spectre expérimental, il est fréquent qu'à la fois un pic et son symétrique soient utilisés, lors de l'alignement, pour aligner deux paquets différents, un sur le pic originel, l'autre sur le complémentaire. Or, dans la pratique, même si cela est possible, ce phénéomène ne se produit pas fréquemment. Aligner à la fois un pic et son symétrique revient à dire qu'il existe dans le spectre un pic qui est issu de la superposition de deux autres pics, le premier représentant un ion C-terminal et le second représentant un ion N-terminal.

Nous avons donc choisi de développer une variante de l'algorithme PacketSpectralAlignment

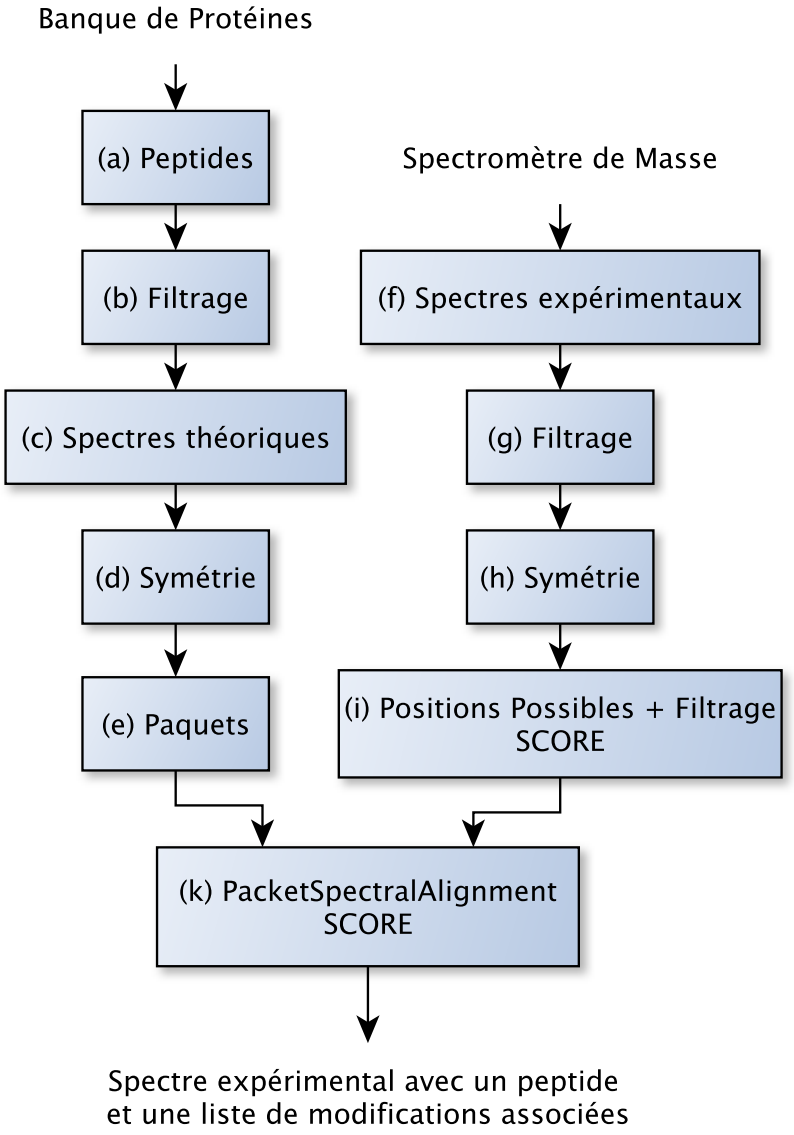


Figure 6.10 – L'enchaînement des composants de la plate-forme SIFpackets.

qui va tenir compte de ce phénomène et le limiter. Puis, après avoir expliqué comment nous utiliserons le score de cette nouvelle méthode de comparaison, nous intégrerons cette variante dans notre plate-forme SIFpackets.

#### 6.3.2.1 Variante avec Liste d'Exclusion

Dans la méthode PacketSpectralAlignment, un pic et son symétrique peuvent tous les deux contribuer au score du peptide. En pratique, comme ils représentent une superposition de pics, il paraît judicieux de ne les prendre en compte dans le score qu'une seule fois et non deux. Cela aura pour principale conséquence d'éviter que la méthode ne crée des modifications pour utiliser ces pics dans les régions pauvres en pics du spectre. Seule la contribution au score est remise en cause, il est donc tout à fait possible de considérer un pic et son symétrique dans l'alignement (la superposition de pics est possible, même si elle reste rare).

Nous proposons ainsi une version modifiée de notre méthode, nommée PacketSpectralAlignment avec liste d'excluions ou **PSAwEL** (pour *PacketSpectralAlignment with Exclusion List*). Dans cette méthode, nous définissons, pour chacune des positions possibles $pp_j$, un ensemble $\mathcal{P}_j$ qui contient tous les pics du spectre expérimental symétrique utilisés quand un paquet est aligné sur $pp_j$. Puis le complémentaire de chaque pic de $\mathcal{P}_j$ est ajouté dans $\mathcal{P}_j$. En connaissant cette information pour l'ensemble des positions possibles du spectre expérimental symétrique, il est possible de définir pour chacune des positions possibles une **liste d'exclusion**. La liste d'exclusion de $pp_j$ est donc un ensemble de positions possibles, qui a pour particularité d'imposer l'utilisation d'un même pic (ou de son complémentaire) lorsqu'elles sont utilisées dans l'alignement. Pour remplir ces listes d'exclusion, une seule règle est utilisée :

La position possible $pp_i$ sera ajoutée à la liste d'exclusion de $pp_j$ si et seulement si l'intersection entre $\mathcal{P}_j$ et $\mathcal{P}_i$ n'est pas vide.

L'idée est donc, durant l'alignement d'un paquet sur une position possible, de vérifier si une position possible de la liste d'exclusion a déjà été utilisée. Si c'est le cas, la contribution au score sera diminuée de sorte à ne pas faire contribuer un pic deux fois. Si la position possible ne figure pas dans la liste d'exclusion, alors la position contribuera normalement au score.

L'algorithme PacketSpectralAlignement est donc modifié tel que présenté dans l'Algorithme 4. Dans cet algorithme, $U(i,j,k)$ représente l'ensemble des positions possibles utilisées pour produire l'alignement $M(i,j,k)$ (défini Section 4.5.5, page 55). La fonction **EstExclus**($U(i,j,k)$, $j'$) permet de savoir si la position possible $j'$ fait partie de la liste d'exclusion d'une des positions possibles de l'ensemble $U(i,j,k)$. Cette fonction va parcourir, pour chacun des éléments de l'ensemble, leur liste de positions possibles exclues. Comme le nombre d'éléments d'une liste d'exclusion n'est dépendant que du nombre de pics du modèle utilisé, et que la taille de l'ensemble est bornée par le nombre d'acides aminés du spectre théorique, la complexité de la fonction est de l'ordre de $O(N)$, où $N$ est le nombre d'acides aminés du spectre théorique.

Nous pouvons noter que dans les cas où un élément de la liste d'exclusion est utilisé dans l'alignement (lignes 6 et 11 de l'Algorithme 4), le score est calculé en tenant compte de ce point. C'est pour cela que la méthode de score utilisée est différente dans ce cas particulier (**scoreExclus**($p_i, pp_j$)).

L'inconvénient majeur de cette modification est de devoir ajouter à chaque étape de l'algorithme une vérification des listes d'exclusions (via la fonction **EstExclus**()). Cela implique, pour chaque étape de l'alignement, de mémoriser les positions possibles utilisées. Or, à chaque

**Algorithm 4** PacketSpectralAlignment avec liste d'exclusion

**Entrée :**
Ensemble des Paquets de $SS_t$ : $\{p_i | \forall i \in [1; N+1]\}$,
Ensemble de Positions Possibles de $SS_e$ : $\{pp_j | \forall j \in [1\,;\, Q]\}$ et
Entier $K$ (nombre maximum de modifications autorisées)

**Sortie :**
Réel $score\_final$ et
Ensemble de modifications $modifications$

**pour** $k$ de 0 à $K$ **faire**
  $D(0,0,k) = 0$
**fin pour**
**pour** $k$ de 0 à $K$ **faire**
  **pour** $j$ de 1 à $Q$ **faire**
    **pour** $i$ de 1 à $N+1$ **faire**
      $(i', j')$ = precurseur$(i, j, k)$
      **si** EstExclus$(U(i', j', k), j)$ **alors**
        $cas1 = D(i', j', k)$ + scoreExclus$(p_i, pp_j)$
      **sinon**
        $cas1 = D(i', j', k)$ + score$(p_i, pp_j)$
      **fin si**
      **si** EstExclus$(U(i-1,$ récupérerPP$(j,$ `PacketSize`$), k-1), j)$ **alors**
        $cas2 = D(i-1,$ récupérerPP$(j,$ `PacketSize`$), k-1)$ + scoreExclus$(p_i, pp_j)$
      **sinon**
        $cas2 = D(i-1,$ récupérerPP$(j,$ `PacketSize`$), k-1)$ + score$(p_i, pp_j)$
      **fin si**
      **si** $cas1 > cas2$ **alors**
        $D(i,j,k) = cas1$
        $U(i,j,k) = U(i', j', k) \cup j$
      **sinon**
        $D(i,j,k) = cas2$
        $U(i,j,k) = U(i-1,$ récupérerPP$(j,$ `PacketSize`$), k-1) \cup j$
      **fin si**
      **si** $M(i-1,j,k) > \max(D(i,j,k), M(i,j-1,k))$ **alors**
        $M(i,j,k) = M(i-1,j,k)$
        $U(i,j,k) = U(i-1,j,k)$
      **sinon**
        **si** $M(i,j-1,k) > \max(D(i,j,k), M(i-1,j,k))$ **alors**
          $M(i,j,k) = M(i,j-1,k)$
          $U(i,j,k) = U(i,j-1,k)$
        **sinon**
          $M(i,j,k) = D(i,j,k)$
        **fin si**
      **fin si**
    **fin pour**
  **fin pour**
**fin pour**

**Algorithm 5** PacketSpectralAlignment avec liste d'exclusion, suite

**pour** $k$ de 0 à $K$ **faire**
  **si** $score\_final < M(N+1, Q, k)$ **alors**
    $meilleur\_k = k$
    $score\_final = M(N+1, Q, k)$
  **fin si**
**fin pour**
$modifications$ = backtrack( $meilleur\_k$ )

étape il y a autant de positions possibles à mémoriser que de paquets alignés, c'est-à-dire $j$. Cela augmente donc la complexité spatiale de $O(NQK)$ à $O(N^2QK)$, où $N$ est le nombre d'acides aminés du peptide représenté par le spectre théorique (donc $N+1$ paquets), $Q$ le nombre de positions possibles du spectre expérimental et $K$ le nombre maximum autorisé de modifications. L'usage de la fonction **EstExclus**() dans l'algorithme d'alignement augmente quant à lui la complexité temporelle de $O(NQK)$ à $O(N^2QK)$.

#### 6.3.2.2 Paramétrage du score de PSAwEL

La version améliorée de notre algorithme nécessite aussi un ajustement de son score. Pour éviter de compter plusieurs fois dans le score un pic se retrouvant aligné deux fois, une règle est ajoutée pour changer le score lors du second alignement. Ce score pourrait être élaboré de différentes manières :

- laissé inchangé, ce qui reviendrait à utiliser la méthode sans liste d'exclusion,
- réduit à 0, ce qui permet de dire qu'un pic ne compte qu'une seule fois,
- ou réduit par un facteur donné, ce qui est un compromis entre les deux solutions précédentes.

En pratique le choix de la solution dépend de plusieurs facteurs importants :

- La superposition de pics est-elle fréquente au sein des spectres ? Dans un tel cas, il serait fréquent de trouver des cas légitimes où un pic devrait contribuer deux fois dans le score. Nous pouvons nous focaliser sur la superposition de pics représentant des ions majeurs (un ion $y$ superposé à un ion $b$), car ce sont les plus importants en terme de score.
- Est-il fréquent d'observer plus d'une superposition dans un même spectre ? En effet, plus le nombre de superpositions dans un même spectre est important, plus leur impact sera fort sur le score de l'alignement.
- Est-ce que ne compter qu'une unique fois un pic dans le score quand il y a effectivement superposition nuit au résultat de l'alignement ? Si cela ne nuit pas, alors réduire le score à 0 lors du second alignement n'est pas problématique.
- Est-ce que le compter deux fois quand il n'y a pas réellement superposition est néfaste pour l'alignement ? Si cela ne nuit pas, alors laisser inchangé le score ne gêne pas à l'identification, et donc la méthode avec liste d'exclusion est inutile.

En pratique, de par le nombre d'acides aminés et leur disparité de masse, les cas de superpositions sont peu fréquents, surtout les cas comportant de nombreuses superpositions. Ces

cas seront généralement formés par les peptides palindromes (le peptide se lit de la même manière de droite à gauche et de gauche à droite). En conséquence de quoi, nous avons choisi l'hypothèse la plus restrictive, n'autoriser la participation au score d'un pic qu'une unique fois, réduisant ainsi la contribution au score de son complémentaire à 0.

#### 6.3.2.3 Intégration de la variante avec liste d'exclusion dans la plate-forme SIFpackets

La variante avec liste d'exclusion de PacketSpectralAlignment, de par sa complexité, va avoir un impact négatif sur le temps d'exécution global de la méthode. Pour cette raison, il ne nous apparaît que peu intéressant de remplacer la méthode PacketSpectralAlignment par sa variante dans la plate-forme. Nous avons donc décidé d'utiliser la première méthode, plus rapide, pour effectuer un criblage de la banque afin de **sélectionner un certain nombre de bons candidats** pour chacun des spectres expérimentaux. Ensuite, nous pourrons ajouter une étape supplémentaire consistant à réévaluer les candidats sélectionnés sur les spectres expérimentaux, à l'aide de la variante plus précise de PacketSpectralAlignment avec liste d'exclusion. Cette réévaluation va introduire le calcul d'un nouveau score, que nous appellerons dans la suite de ce document le **score ajusté**. La Figure 6.11 présente la plate-forme SIFpackets dans laquelle a été intégrée la variante avec liste d'exclusion de l'algorithme de comparaison.

**Sélection des candidats pour une évaluation ajustée du score.**

Il existe de nombreuses possibilités pour sélectionner les candidats dont on souhaite réévaluer le score. Parmi ces possibilités, nous pouvons citer :

- Garder les $X$ meilleurs résultats fournis par PacketSpectralAlignment, $X$ étant un entier constant fixé au préalable par l'utilisateur. L'avantage est que cela permet une parfaite maîtrise du temps passé à recalculer les scores des candidats, ce temps ne dépendant plus que du nombre de spectres à traiter.

- Conserver un nombre de candidats proportionnel à la taille de la banque, ou plus précisément proportionnel à la partie de la banque considérée lors de la comparaison, c'est-à-dire en tenant compte des filtrages éventuels.

- Fixer un seuil et ne conserver que les candidats dont le score a dépassé ce seuil. En théorie, cela permettrait d'éliminer les spectres pour lesquels l'identification correcte ne se trouve pas dans la banque, en supposant que ceux-ci ont un score inférieur aux identifications correctes. En pratique, il y a une variabilité importante des scores d'un spectre à l'autre, due à différents facteurs tels que le bruit, ou simplement la composition du peptide qui va altérer la fragmentation et donc le score.

Dans la solution que nous avons retenue, nous tenons compte du fait que le score fourni par PSAwEL est inférieur ou égal au score fourni par PacketSpectralAlignment. Cela vient du fait que la méthode avec liste d'exclusion cherche une solution dans un sous-ensemble des solutions de PacketSpectralAlignment. Dans les cas où l'on n'utilise pas un pic et son complémentaire en même temps, le score restera identique, tandis que dans les cas contraires, le score sera diminué. Il est donc possible, comme décrit dans l'Algorithme 6, de sélectionner le meilleur

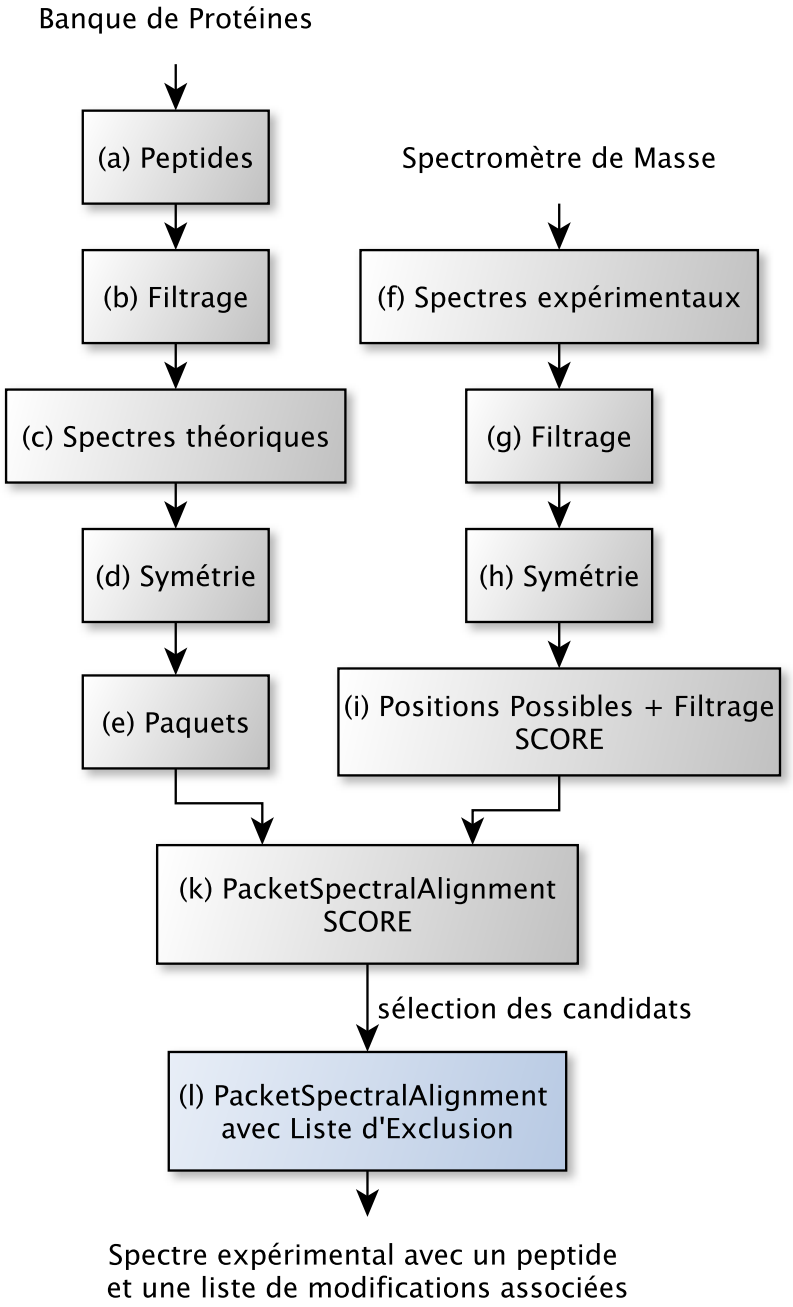


Figure 6.11 – L'enchaînement des composants de la plate-forme SIFpackets avec l'ajout de la variante avec liste d'exclusion.

candidat en se basant sur le score de PacketSpectralAlignment, puis d'utiliser le score de la

variante avec liste d'exclusion (PSAwEL) pour filtrer les autres candidats. En effet, si l'ancien score n'est pas supérieur au meilleur nouveau score obtenu jusque là, alors il est inutile de le réévaluer à l'aide de la variante. Plus la première méthode de comparaison (ici en l'occurrence PacketSpectralAlignment) donne de bons résultats, moins il y aura de candidats à réévaluer.

Il faut cependant prendre garde à un cas particulier, celui où le peptide correspondant au spectre n'est pas dans la banque. Dans un tel cas, les meilleurs candidats n'auront pas nécessairement un score très élevé, et seront certainement nombreux à avoir des scores semblables. Dans ce cas, cela pourrait impliquer de devoir réévaluer de très nombreux peptides. Il est donc intéressant d'ajouter une borne supérieure au nombre de candidats à réévaluer, pour limiter un potentiel impact négatif sur le temps d'exécution.

**Algorithm 6** Algorithme de calcul du score ajusté

**Entrée :**
$S(s, p)$ tableau contenant le score de PSA pour l'association du spectre $s$ avec le peptide $p$
$B$ ensemble de tous les peptides composant la banque
$E$ ensemble de tous les spectres expérimentaux

**Sortie :**
$S'(s, p)$ tableau réduit de scores ne contenant que les meilleurs scores calculés à l'aide de PSAwEL

```
1: pour tout spectre expérimental SS_e ∈ E faire
2:   score_arret = 0
3:   pour tout peptide p_i ∈ B faire
4:     si S(SS_e, p_i) >= score_arret alors
5:       S'(SS_e, p_i) = PSAwEL(SS_e, p_i)
6:       score_arret = max( score_arret, S'(SS_e, p_i) )
7:     fin si
8:   fin pour
9: fin pour
```

**Utilisation du score ajusté dans l'identification des peptides.**

Une fois le score ajusté calculé pour les candidats sélectionnés à l'aide de l'algorithme PacketSpectralAlignment avec Liste d'Exclusion, il faut tenir compte de ce nouveau score. Nous pourrions l'utiliser pour consolider les résultats. Ainsi, si un même candidat obtient le meilleur score avec les deux calculs de score, il peut être jugé comme fiable. Dans le cas contraire, il est par exemple possible d'ignorer le spectre et donc de perdre une identification potentielle. L'avantage de cette solution est qu'elle offre une certitude assez forte sur les identifications, tout en permettant d'éliminer des faux positifs, l'inconvénient étant qu'elle interdit toute amélioration des résultats, en refusant à la réévaluation de changer l'ordre des candidats.

Nous pouvons également utiliser le score ajusté pour reclasser les peptides candidats en bénéficiant pleinement des capacités de PacketSpectralAlignment avec Liste d'Exclusion.

Nous avons choisi ici de retenir cette dernière solution. Il est à noter que pour chacun des candidats, nous conservons à la fois le score avant et après sa réévaluation. Ces scores pourront

être utilisés ultérieurement, notamment lors de la remontée à la protéine. Cela va permettre ainsi d'utiliser la différence entre les deux scores comme une mesure de confiance, comme nous le verrons dans la Section 6.4.

### 6.3.3 Résultats expérimentaux : ISB Dataset

Les premiers tests effectués visent à comparer l'algorithme PacketSpectralAlignment avec l'algorithme SpectralAlignment sur le jeu de données spectres_ISB décrit dans la Section 5.2.1.1, page 62. Cette comparaison permet d'évaluer le comportement des méthodes en présence de modifications. Pour cela, nous avons modifié la banque contenant les protéines à identifier comme décrit dans la Section 5.2.2.3, page 66.

Afin d'évaluer les résultats fournis par les deux méthodes sur les données modifiées de l'ISB, nous devons fixer la valeur du paramètre $D$ qui sert à définir ce qu'est un résultat attendu (voir Section 5.3.2, page 67). Nous avons défini $D$ de manière semblable au nombre de modifications recherchées par les méthodes, c'est-à-dire :

$$D = ((M_{peptide}/0,0015) + 1)$$

Dans cette équation $D$ est donc identique au calcul du nombre de modifications autorisées (paramètre $K$) au $+1$ près. L'ajout du facteur 1 intervient ici car, comme nous l'avons précédemment expliqué, les algorithmes testés autorisent une modification de type re-calibrage en début de spectre, modification qui n'intervient pas dans le paramètre $K$.

La notion de résultat attendu permet de dire, pour chaque résultat produit par une des méthodes de comparaison, s'il s'agit d'un vrai positif (identification réussie d'un résultat attendu), un faux positif (identification positive d'un résultat non attendu), un vrai négatif (identification négative d'un résultat non attendu) ou un faux négatif (identification négative d'un résultat attendu). Avec ces données, il est possible de tracer la courbe ROC et de calculer l'aire sous cette courbe (AUC).

La Figure 6.12 indique les différentes AUC obtenues (a) par la méthode PacketSpectralAlignment (courbe bleue) (b) par la méthode SpectralAlignment (courbe rouge), dans les mêmes conditions, et pour différents niveaux de modifications. Nous pouvons noter que les résultats fournis par PacketSpectralAlignment sont nettement supérieurs, quel que soit le nombre de modifications présentes. Nous pouvons cependant noter que pour les deux méthodes, les résultats se dégradent sensiblement lorsque la banque est modifiée à l'aide d'une matrice PAM60 ou PAM80 (ce qui correspond en moyenne à moins de 60% d'identité entre les séquences modifiées et les séquences non modifiées).

L'observation des courbes ROC de la Figure 6.13 permet de remarquer une pente beaucoup plus importante de la courbe représentant la méthode PacketSpectralAlignment pour un faible taux de faux positifs. Ainsi, si nous fixons le taux de faux positifs à 1%, comme c'est fréquemment le cas lors des analyses, dans le cas de la banque modifiée à l'aide d'une matrice PAM10, nous obtenons 47,1% de vrais positifs avec la méthode PacketSpectralAlignment et seulement 13,5% avec SpectralAlignment. L'écart de résultat entre les deux méthodes est donc particulièrement accentué pour un faible taux de faux positifs (jusqu'à 10% environ), ce qui est tout à fait

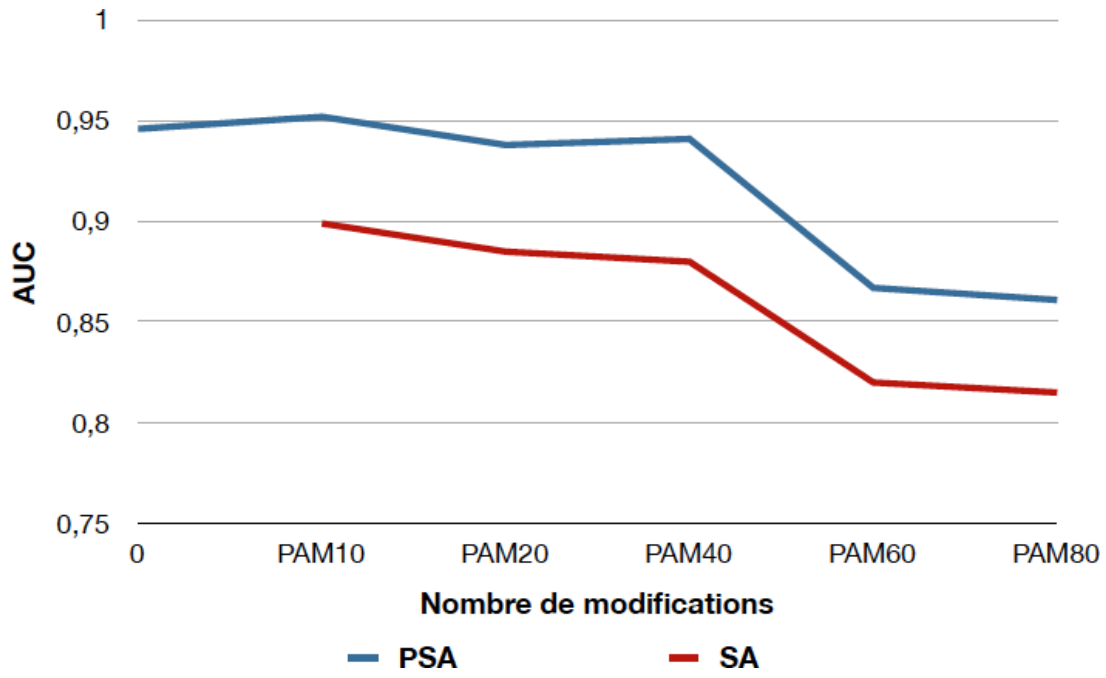


Figure 6.12 – Évaluation de la qualité (AUC) des résultats des méthodes PacketSpectralAlignment (PSA) et SpectralAlignment (SA) en fonction du niveau de modification de la banque.

intéressant et s'atténue par la suite.

Le Tableau 6.3 indique le temps moyen passé par chacune des méthodes pour analyser les données spectres_ISB, ainsi que le temps moyen pour analyser un spectre avec la banque 18mix_rice1700. Nous voyons immédiatement que le temps d'exécution de PSA, grâce eux différents filtres liés aux positions possibles ainsi qu'à sa notion de paquet, a pu être divisé par cinq, ce qui représente une amélioration très significative.

| Méthode | Temps moyen (spectres_ISB) | Temps moyen (par spectre) |
|---|---|---|
| PacketSpectralAlignment | 16.800 secondes | 28,5 secondes |
| SpectralAlignment | 84.000 secondes | 142,5 secondes |

Table 6.3 – Comparaison des temps d'exécution de PacketSpectralAlignment et de SpectralAligmment sur le jeu de données spectres_ISB.

### 6.3.4 Résultats expérimentaux : Brachypodium

Nous avons poursuivi les tests en évaluant la méthode PacketSpectralAlignment avec liste d'exclusion pour évaluer l'intérêt des scores ajustés, c'est-à-dire l'impact de l'étape (l) de la plateforme (Figure 6.10, page 85). Nous avons donc analysé une première fois un jeu de spectres en

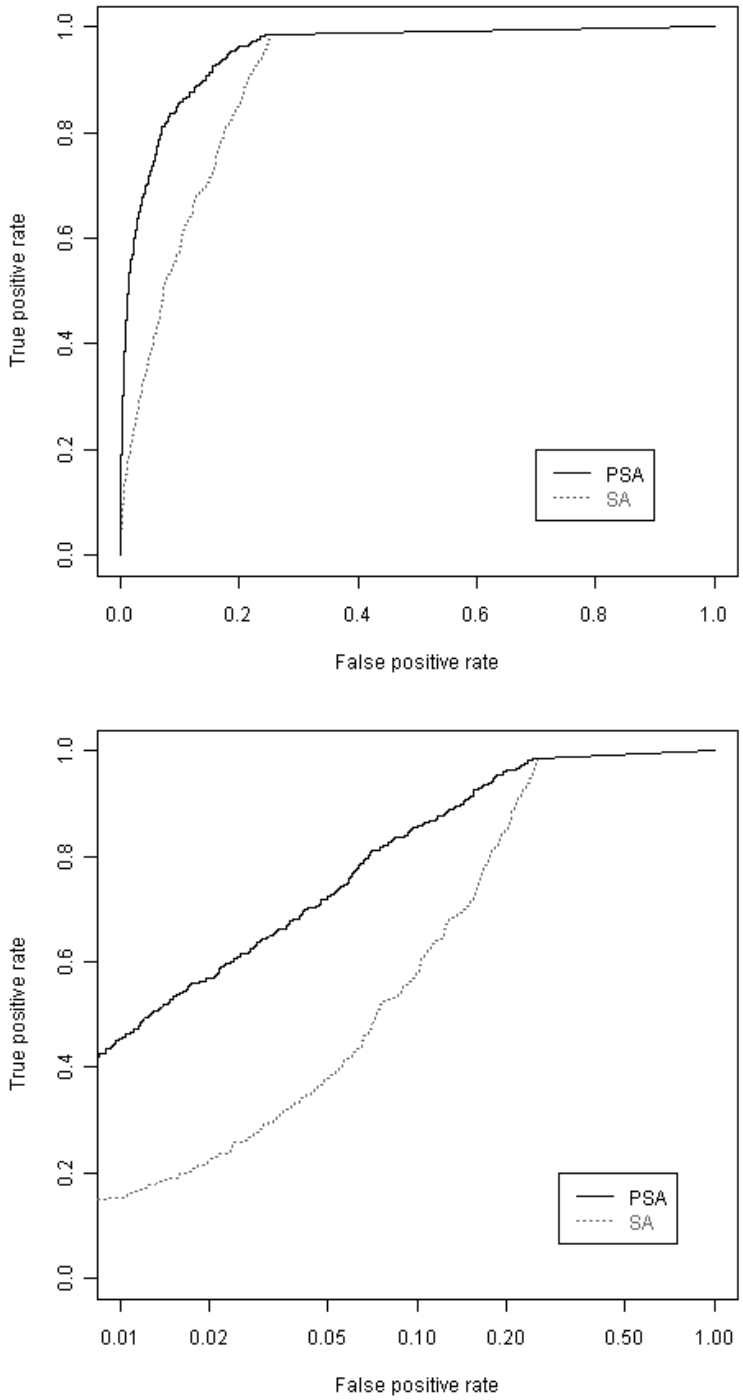


Figure 6.13 – Courbes ROC représentant les résultats de PacketSpectralAlignement (en noir) et de SpectralAlignment (en gris). Les deux figures représentent les mêmes données, mais la figure du bas est tracée en échelle logarithmique, ce qui permet de mettre en évidence la différence de comportement pour les faibles taux de faux positifs.

n'appliquant pas l'étape (l), puis une seconde fois en l'appliquant, et nous avons comparé le nombre de peptides identifiés dans les deux cas.

Cette comparaison a été effectuée sur le jeu de données Brachypodium **bradi_D1** à **bradi_D8** défini Section 5.2.1.2, page 63, lequel contient de nombreux spectres correspondant à des peptides non modifiés, ainsi que quelques spectres présentant des modifications post-traductionnelles. Pour ce jeu de données, nous connaissons la protéine identifiée correspondant à chaque spot. Notre objectif sera donc de trouver un maximum de peptides appartenant à cette protéine, chacun de ces peptides obtenant un score supérieur à n'importe quel autre peptide de la banque. Ces comparaisons ont été effectuées en utilisant la banque **Bradi_1g**, défini Section 5.2.2.2, page 65.

| | **Spot 1** | **Spot 2** | **Spot 3** | **Spot 4** | **Spot 5** | **Spot 6** | **Spot 7** | **Spot 8** |
|---|---|---|---|---|---|---|---|---|
| **Nb de Spectres** | 291 | 249 | 240 | 122 | 146 | 357 | 143 | 91 |
| **Nb de peptides identifiés sans score ajusté** | 7 | 28 | 18 | 2 | 7 | 46 | 7 | 2 |
| **Nb de peptides identifiés avec score ajusté** | 13 | 36 | 30 | 2 | 9 | 67 | 7 | 3 |
| **Nb de peptides différents identifiés sans score ajusté** | 5 | 8 | 7 | 2 | 5 | 10 | 4 | 1 |
| **Nb de peptides différents identifiés avec score ajusté** | 9 | 12 | 12 | 2 | 6 | 13 | 5 | 2 |
| **Nb de modifications localisées** | 3 | 10 | 8 | 1 | 2 | 6 | 3 | 0 |

Table 6.4 – Nombre de peptides identifiés lors de la comparaison des spectres de Brachypodium (**bradi_D1** à **bradi_D8**) sur la banque **Bradi_1g** en utilisant dans un premier temps PacketSpectralAlignment sans l'option de score ajusté, puis dans un second temps avec le score ajusté.

Les résultats présentés dans le Tableau 6.4 mettent en évidence l'apport de la variante de l'algorithme avec liste d'exclusion. Nous pouvons observer que le nouveau calcul des scores augmente de $50\%$ le nombre de peptides identifiés sur ce jeu de données, et même de plus de $50\%$ le nombre de peptides différents identifiés. Augmenter le nombre de peptides identifiés signifie trouver plus de spectres identifiant un même peptide, et donc une certitude plus importante quant à la présence du peptide dans l'échantillon, tandis que l'augmentation du nombre de peptides différents identifiés va permettre une meilleure discrimination entre les protéines, en augmentant la couverture de séquences identifiées sur celle-ci.

La dernière ligne du Tableau 6.4 indique le nombre de peptides identifiés comportant une modification post-traductionnelle. Nous avons pu vérifier la majorité de ces modifications en nous comparant aux résultats donnés par une analyse à l'aide du logiciel Mascot. En effet, la

plupart de ces modifications était possible à anticiper par ce logiciel, étant issues principalement des protocoles appliqués à l'échantillon. Mais outre ces modifications, nous avons pu déceler des modifications post-traductionnelles non identifiées par Mascot, telle que celle présentée dans la Figure 6.14. Dans cette figure, le spectre du bas est le spectre théorique symétrique du peptide QQQQQEGEEEGFIIR, tandis que le spectre du haut est un spectre expérimental issu de Brachypodium. Une comparaison de ces deux spectres fait apparaître une modification de +28 daltons sur le premier acide glutamique (E). Cette modification post-traductionnelle est très probablement liée à la coloration des spots au bleu de Coomassie et a été signalée assez récemment [JLW+08, SB09].

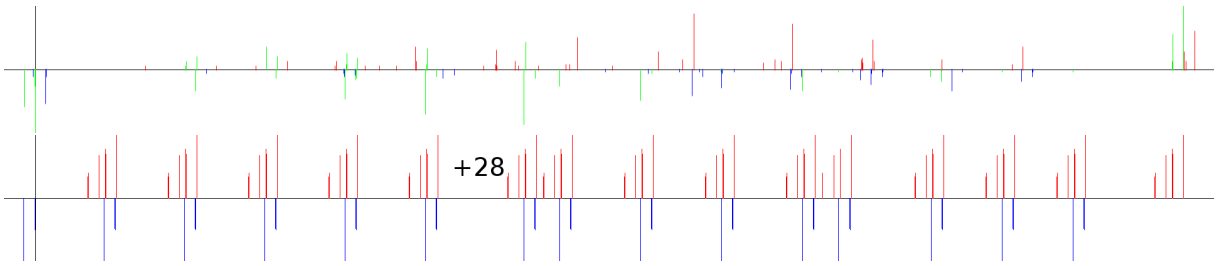


Figure 6.14 – Alignement d'un spectre expérimental comportant une modification identifiée et localisée à l'aide de la plate-forme SIFpackets -spectre du haut- avec un spectre théorique -spectre du bas-.

## 6.4 Remontée à la protéine

À partir des résultats obtenus par analyse MS/MS, le chercheur souhaite en général retrouver les protéines contenues dans un échantillon. Nous pouvons remarquer que le terme d'identification, dans le cas des organismes non séquencés, est abusif puisque la protéine réelle peut ne pas être présente dans la banque de référence, mais que l'on espère retrouver ses protéines homologues.

Le résultat final proposé à l'issue d'une interprétation donnée se présente ainsi sous la forme de la liste des protéines les plus probables. L'identification des peptides à partir des spectres MS/MS est donc une étape essentielle mais intermédiaire pour extrapoler ces protéines.

Cependant, l'identification d'une protéine à partir d'un ensemble de peptides est un problème qui est loin d'être trivial, et qui a été présenté sous le terme de "protein inference problem" par Nesvizhski en 2005 [NA05]. Le problème est particulièrement complexe dans les approches de type "shotgun" présentées Section 2.3.4.2, page 21, car suite à une étape de digestion des protéines très en amont du processus d'analyse, le lien entre les peptides et la protéine d'origine est perdu. Or, surtout dans le cas des organismes eucaryotes, la banque de référence va souvent contenir de nombreuses séquences homologues, difficiles à distinguer les unes des autres.

Nous devons compléter la plate-forme SIFpackets pour présenter la liste des protéines les plus probables à l'utilisateur. Pour cela, nous allons attribuer un score à chaque protéine, les protéines associées au plus grand score étant les plus probables.

### 6.4.1 Établir un score pour les protéines

La multiplication des identifications incorrectes de protéines a conduit les éditeurs de journaux à énoncer en 2004 un ensemble de règles à respecter. Ces règles sont regroupées dans un document intitulé “The need for guidelines in publication of peptide and protein identification data” [CAB+04].

Parmi les critères de fiabilité, la couverture de la séquence protéique par les peptides identifiés est considérée par la communauté comme un élément de confiance important. Cependant, le nombre de peptides “identifiables” par MS/MS peut varier fortement d'une protéine à l'autre, en fonction de la taille de la protéine et des propriétés physico-chimiques des peptides.

Nous avons choisi de procéder au calcul des scores associés aux protéines de la banque en deux étapes distinctes :

- Une première étape donne un score à chacun des peptides identifiés afin d'évaluer leur apport en terme d'identification de protéine,
- une seconde étape combine les scores des différents peptides pour définir le score d'une protéine.

#### 6.4.1.1 Score d'un peptide : importance d'un peptide dans une identification

Nous avons choisi de définir, pour chaque peptide, un score reflétant l'importance de l'identification de ce peptide dans l'identification d'une protéine. En effet, deux peptides différents n'apporteront pas nécessairement la même certitude quant à l'identification d'une protéine. Ainsi, nous avons choisi d'utiliser différents paramètres pour calculer ce score :

- La longueur d'un peptide en terme de nombre d'acides aminés. En effet, plus un peptide est long, plus il apportera une certitude forte à l'identification, car il y a moins de chance de se tromper sur l'interprétation du spectre, et qu'il apportera une couverture plus importante sur la protéine.
- Le nombre de protéines de la banque qui partagent le peptide. Car, plus un peptide est présent dans un nombre important de protéines, moins il sera intéressant pour l'identification.
- La complexité en acides aminés du peptide. Certaines parties des protéines sont en effet dites **à faible complexité** [HG10]. Ces zones sont généralement constituées d'une forte répétition d'un petit nombre d'acides aminés, ces acides aminés étant généralement ceux de plus petite masse. Il en ressort généralement des peptides peu utiles à l'identification, car il produisent des spectres qui pourront être associés aisément à de nombreux autres peptides à faible complexité.
- La confiance que l'on a en l'association spectre-peptide. Nous pouvons, en utilisant une combinaison des scores produits par PacketSpectralAlignment et par PacketSpectralAlignment avec Liste d'Exclusion, définir un indice de confiance.

L'Algorithme 7 permet de calculer le score de chacun des peptides. Cet algorithme prend en entrée un peptide, le nombre de protéines de la banque contenant ce peptide, ainsi que les scores de comparaison associés à ce peptide avant et après le calcul du score ajusté décrit Section 6.3.2. La fonction **estDeFaibleComplexité**(p) permet de déterminer si le peptide p est considéré comme ayant une faible complexité. Pour déterminer si un peptide est de faible com-

plexité, il existe des outils comme SEG [WF96]. La constante PÉNALITÉ_COMPLEXITÉ est ici une valeur supérieure à 1, qui va permettre de diminuer l'importance accordée aux peptides de faible complexité. Nous avons choisi de fixer cette constante à 3 lors de nos tests. Enfin, la fonction **longueur**(p) donne la longueur en nombre d'acides aminés du peptide p.

**Algorithm 7** Algorithme de calcul du score d'un peptide lors de la remontée à la protéine.

**Entrée :**
Chaîne peptide
Entier nb_proteines
Réel $S_{PSA}$
Réel $S_{PSAwEL}$

**Sortie :**
Réel score

1: pénalité = 1
2: **si** estDeFaibleComplexité(peptide) **alors**
3: pénalité = PÉNALITÉ_COMPLEXITÉ
4: **fin si**
5: confiance = $S_{PSAwEL}$ / $S_{PSA}$
6: score = $\frac{\text{longueur(peptide) * confiance}}{\text{nb\_protéines * pénalité}}$
7: **return** score

#### 6.4.1.2 Score de la protéine

Une fois un score calculé pour chacun des peptides, il est nécessaire d'établir un score pour une protéine donnée à partir des scores des peptides appartenant à cette protéine. Nous avons choisi de considérer les scores des peptides de deux manières différentes :

- Tout d'abord, faire contribuer au score d'une protéine donnée absolument toutes les identifications correspondant à un de ses peptides. Ainsi, si un même peptide a été identifié plusieurs fois (à l'aide de différents spectres), il contribuera plus fortement à l'identification. Il paraît en effet raisonnable de renforcer l'hypothèse de présence d'un peptide qui a été identifié plusieurs fois par différents spectres.
- Ensuite, nous considérons la contribution des peptides distincts les uns des autres. Cela permet de faire ressortir qu'une identification avec des peptides distincts se traduira par une meilleure couverture de la protéine.

Le score de la protéine est calculé de la sorte :

$$S_P = \left( \frac{cumul * 100}{maxCumul} + \frac{cumulDistinct * 100}{maxCumulDistinct} \right) /2 \tag{6.1}$$

Où :

- $cumul$ est la somme des scores $S$ de tous les peptides identifiés dans l'analyse appartenant à la protéine. Ainsi, 10 spectres donnant un unique peptide ajouteront 10 fois le score

$S$ au $cumul$. $maxCumul$ contiendra le plus gros cumul observé pour une protéine de la banque.
– $cumulDistinct$ est la somme des scores S de tous les peptides identifiés associés à la protéine, moins les doublons. Ainsi, 10 spectres donnant un unique peptide n'ajouteront qu'une fois le score $S$ au $cumulDistinct$. $maxCumulDistinct$ contiendra le plus gros $cumulDistinct$ observé pour une protéine de la banque.

Lorsque chacune des protéines de la banque s'est vue attribuer un score, il est possible de les ordonner par score $S_P$ décroissant. Le score le plus élevé correspond à la meilleure identification.

### 6.4.2 Résultat sur les données de Brachypodium

Nous avons pu conduire différents tests de remontée à la protéine sur les huit jeux de données **bradi_D1** à **bradi_D8**. Nous avons cherché à évaluer la capacité de notre score à identifier la protéine escomptée (en utilisant le rang de cette protéine), ainsi qu'à la discriminer des autres protéines (en évaluant le delta de score la séparant de la meilleure protéine non recherchée). Les résultats de ces tests sont présentés dans le Tableau 6.5.

Nous pouvons observer que le score que nous avons développé a amélioré l'identification de la protéine attendue, et ce surtout sur les jeux de spectres de pauvre qualité. Il est important de noter ici que le delta séparant une identification correcte de rang 1 de la seconde identification (incorrecte a priori) est assez important, et montre donc que ce score discrimine plutôt bien les résultats. Nous pouvons faire une remarque complémentaire sur le jeu de spectres **bradi_D4** qui est de très mauvaise qualité. Sur ce jeu de spectres, l'identification est quasiment impossible, et ce quel que soit le logiciel utilisé. Nous voyons ici que la protéine escomptée n'est pas mal classée, mais a un score insuffisant. Nous pouvons aussi souligner que, dans ce cas précis, le delta entre la protéine de rang 1 et celle de rang 2 est plutôt faible (37.34) en comparaison de ce que nous pouvons observer sur les autres spots, ce qui montre qu'il y a une incertitude forte quant à la fiabilité de l'identification. Enfin, le dernier point important dans le score est la normalisation. Lorsque le cumul des scores des peptides est utilisé, il n'est pas possible, en n'ayant qu'un score, de dire si il s'agit ou non d'une identification, ni même si le score est "bon" ou non. En revanche, avec le score de SIFpackets, le score étant normalisé, nous savons qu'un bon score est proche de 100.

| | **cumul score peptide** | | **score SIFpackets** | | |
|---|---|---|---|---|---|
| | **score (delta)** | **score distinct (delta)** | **score** | **rang** | **delta** |
| **bradi_D1** | 134 (-3) | 90 (+17) | 100.00 | 1 | 61.68 |
| **bradi_D2** | 495 (+393) | 131 (+86) | 100.00 | 1 | 87.46 |
| **bradi_D3** | 307 (+239) | 117 (+64) | 100.00 | 1 | 79.53 |
| **bradi_D4** | 45 (-7) | 45 (-7) | 46.66 | 5 | -53.34 |
| **bradi_D5** | 118 (+16) | 81 (+44) | 100.00 | 1 | 71.24 |
| **bradi_D6** | 562 (+492) | 138 (+108) | 100.00 | 1 | 89.22 |
| **bradi_D7** | 116 (+74) | 65 (+30) | 100.00 | 1 | 56.82 |
| **bradi_D8** | 46 (+4) | 23 (-2) | 98.00 | 1 | 44.62 |

Table 6.5 – Comparaison des scores des protéines sur les données **bradi_D1** à **bradi_D8**.

# CHAPITRE 7

# Conclusions et perspectives

## 7.1 Conclusions

Dans ce mémoire, nous nous sommes intéressés à l'identification des protéines dans le cas d'organismes non séquencés. Nos travaux se sont surtout orientés vers la création d'une méthode permettant d'associer des peptides à des spectres MS/MS, tout en tolérant des modifications dans la séquence peptidique. Afin d'effectuer cette association, nous avons choisi de nous orienter vers une approche dite de comparaison de spectres et de porter un soin particulier au respect de la structure propre des spectres MS/MS. Notre méthode utilise donc de nombreuses informations relatives à la manière dont sont construits les spectres.

Nous avons tout d'abord développé l'algorithme PacketSpectralAlignment [CFRT09]. Il s'agit d'une approche basée sur la programmation dynamique qui permet de trouver le meilleur alignement possible existant entre un spectre expérimental MS/MS et un spectre théorique créé à partir d'une séquence peptidique. Notre méthode a pour point fort de respecter la construction des spectres, la logique avec laquelle les spectromètres les créent, ce qui est rendu possible via l'usage des notions de symétrie et de paquets. Notre méthode est capable, grâce à ces deux notions, de comparer des spectres même lorsque de nombreuses modifications sont présentes. En effet, comme nous l'avons montré (Figure 3.11, page 43), une représentation ne respectant pas les contraintes de construction d'un spectre nuit à la prise en compte des modifications lors de l'alignement.

Nous avons ensuite développé des stratégies visant à améliorer les performances, que ce soit en terme de qualité ou de temps d'exécution, de l'algorithme PacketSpectralAlignment. Ces stratégies consistent en différentes méthodes de filtrage opérant à différentes étapes de la comparaison. Cela va d'un filtrage sur les données de la banque à un filtrage des spectres. Ces filtrages ont été développés et paramétrés de sorte à tenir compte au mieux du fonctionnement de PacketSpectralAlignment via la prise en compte de la symétrie et l'utilisation des paquets. Nous avons détaillé le paramétrage de chacune de ces méthodes et les choix effectués ; nous avons aussi détaillé la manière de paramétrer notre algorithme d'alignement. Tous ces éléments ont été intégrés dans une plate-forme que nous avons nommé SIFpackets [CFRT10].

Enfin, nous avons pu évaluer notre plate-forme SIFpackets sur différents jeux de données.

Dans un premier temps, l'évaluation de SIFpackets a été menée sur des données parfaitement connues, et nous avons pu (1) donner des mesures exactes quant à la qualité de nos résultats

et (2) nous comparer à la méthode SpectralAlignment [PDT00]. Cette comparaison a permis de mettre en avant un gain important pour notre méthode. Ce gain est visible en termes de qualité : nous avons amélioré d'un facteur $3,5$ le taux d'identifications correctes (vrais positifs) lorsque nous autorisons un maximum de $1\%$ d'identifications erronées (faux positifs) ; mais le gain est aussi visible en termes de vitesse d'exécution : notre algorithme est en effet $5$ fois plus rapide que SpectralAlignment (résultats visibles Section 6.3.3, page 92).

Puis, dans un second temps, nous avons évalué SIFpackets sur un jeu de données où seule la protéine escomptée était connue. Ce jeu nous a permis de montrer l'apport d'une variante de notre algorithme, appelée PacketSpectralAlignment avec liste d'exclusion (PSAwEL). Cette variante permet de mieux localiser les modifications au sein des spectres MS/MS, comme les résultats ont permis de le vérifier. Ainsi, sur ce jeu de données, utiliser PSAwEL permet une augmentation de $50\%$ du nombre de peptides identifiés. Nous avons aussi pu, au cours de ces tests, identifier des modifications post-traductionnelles qui n'étaient pas attendues.

## 7.2 Perspectives

Nous avons présenté ici notre méthode PacketSpectralAlignement qui, intégrée dans notre plate-forme SIFpackets, permet d'associer des peptides à des spectres MS/MS, et ce même en présence de nombreuses modifications. Notre méthode, comme nous avons pu le montrer, donne des résultats convaincants. Cependant, elle pourrait gagner à utiliser d'autres informations, ou à être combinée avec d'autres techniques déjà existantes afin d'en améliorer la précision et d'en réduire davantage le temps d'exécution.

Nous proposons ici un ensemble d'idées qui restent à évaluer en différents points du processus d'identification des spectres MS/MS pour éventuellement améliorer l'identification.

**Modification du protocole expérimental.** Une des premières possibilités s'offrant à nous pour améliorer l'identification est de modifier certains aspects du protocole expérimental. Ainsi, il a déjà été proposé par Bandeira et al [BCP07] de combiner les analyses issues d'un même mélange hydrolysé par différentes enzymes. Cette approche permet tout d'abord d'identifier de nouveaux peptides dont la masse se serait trouvée hors des capacités d'analyse d'un spectromètre de masse avec l'utilisation d'une seule enzyme. Elle permet également d'utiliser l'information de chevauchement entre les peptides générés de manière similaire à ce qui est couramment effectué dans le séquençage de génomes avec la méthode dite de *whole-genome shotgun* [Sta79, FAW+95]. Cette méthode peut s'avérer intéressante, mais a cependant un coût en terme de préparation d'échantillons et de temps d'analyse. De plus, il faut disposer de moyens différents pour hydrolyser efficacement les protéines à analyser, ce qui n'est pas toujours applicable, en raison, notamment, de la composition en acides aminée des protéines.

**Usage des données issues de l'étape de chromatographie.** La chromatographie en phase liquide est une étape de séparation des peptides précédant fréquemment l'analyse en spectrométrie de masse. Le temps de rétention qui représente le temps passé par le peptide dans la chromatographie est une ifnormation qui est en grande partie ignorée, alors qu'elle pourrait

être très utile. En effet, il a déjà été montré qu'il était possible d'identifier des protéines en utilisant uniquement le temps de rétention [PRM+02]. Il pourrait être intéressant d'associer cette information à notre processus d'identificaiton dans le cas où des modifications sont présentes. Nous devons cependant souligner que la chromatographie en phase liquide n'est pas la seule méthode de séparation utilisée en spectrométrie de masse. Développer une telle approche pour chacune des techniques de séparation n'est probablement pas possible, et impliquerait des solutions potentiellement très différentes les unes des autres.

**Quantification des peptides.** Les spectromètres de masse fournissent de plus en plus de données quantitatives quant à la présence des peptides dans un échantillon. Utiliser de telles données apporterait une information supplémentaire lors de l'identification, et pourrait permettre de discriminer les protéines dans certains cas. Cependant, même si dans l'échantillon, nous pouvons nous attendre à observer une quantité similaire de toues les peptides appartenant à une même protéine, après analyse en spectrométrie de masse, on observe des fréquences très différentiées dues à des efficacités d'ionisation de peptides très variables. La prise en compte de la fréquence des peptides est donc délicate.

**Filtrage par Tags.** À l'instar de nombreuses autres méthodes de comparaison de spectres, utiliser un module de filtrage à partir des tags pourrait permettre de réduire considérablement l'espace de recherche. Ces méthodes, générant de courtes séquences d'acides aminés en appliquant une approche de type *de novo* sur un spectre, présentent pour avantage d'être relativement rapides et efficaces, comme expliqué dans le Chapitre 3, page 38. Les approches par tags existantes sont nombreuses et habituellement satisfaisantes, mais une adaptation au cadre des organismes non séquencés est nécessaire. En effet, contrairement au cas des organismes séquencés où très peu de modifications sont attendues, nous nous attendons à en trouver un nombre plus conséquent, certainement plusieurs par peptide. Cela signifie qu'il faut prendre en compte le risque qu'une modification ait eu lieu sur un tag. Un tag trop long augmente donc les chances qu'il contienne des modifications. De plus, du fait que les protéines peuvent contenir de nombreuses modifications, au point parfois d'en faire apparaître ou disparaître de longs morceaux de séquence, il faut considérer qu'un unique tag n'est pas suffisant pour filtrer les protéines hors des candidats.

Le développement d'une interface utilisateur très ergonomique faciliterait l'utilisation en routine de la plateforme SIFpackets dans les laboratoires d'analyse de spectrométrie de masse. En particulier, une méthode telle que celle développée dans ce mémoire gagnerait fortement à être couplée à une application visuelle d'annotation des spectres, c'est-à-dire une application permettant de visualiser des spectres MS/MS sur lesquels les acides aminés identifiés et les éventuelles modifications seraient indiqués. Un tel outil pourrait permettre au biologiste de valider les identifications et modifications trouvées, tout en offrant éventuellement à celui-ci la possibilité d'agir sur cette identification. Nous pouvons en effet imaginer des cas où différentes modifications sont envisageables, où l'algorithme en choisit une en particulier, mais où le biologiste, avec les connaissances supplémentaires dont il dispose, pourrait en choisir une autre. Dans tous les cas de figure, un tel outil, couplé à notre algorithme, permettrait de réduire

considérablement la tâche du biologiste qui doit encore aujourd'hui, dans de très nombreux cas, interpréter et annoter les spectres manuellement, une tâche longue et fastidieuse.

Nous pouvons aussi souligner que fournir des spectres annotés est absolument nécessaire pour publier des résultats dans les meilleures revues du domaine lorsque les spectres portent certaines modifications post-traductionnelles.

L'outil d'annotation des spectres pourrait aussi chercher à interpréter les modifications, ou du moins aider le biologiste à le faire en fournissant une liste de modifications possibles. Pour ce faire, une banque répertoriant les modifications existantes pourrait être utilisée. Il s'agit d'une étape qui peut être très utile, mais que nous avons délibérément choisi de ne pas effectuer dans notre méthode, afin d'autoriser la découverte de modifications jusque là inconnues.

Si nous avons ici brièvement abordé le problème de la remonté à la protéine (Section 6.4, page 96), notre solution nécessiterait certainement des travaux supplémentaires pour être efficace dans le cadre de mélanges complexes de protéines. Le problème de l'identification des protéines en mélange décrit par [NA05] est déjà très difficile dans le cas des organismes séquencés, pourtant plus simple que celui des organismes non séquencés.

Par ailleurs une des difficultés majeures, qui n'est pas encore résolue dans le cas des identifications avec modification est la validation automatique des résultats. Habituellement, des banques de type "decoy" servent à évaluer le taux de faux positifs, mais cette méthode d'évaluation est inadaptée lorsque l'on autorise des modifications [AML10]. Ainsi, à l'heure actuelle, aucune méthode ne permet de valider automatiquement des identifications dans el cas où des modifications sont tolérées. Malgré les avancées de ces dernières années par de nombreuses équipes et l'apport de notre travail, il reste encore de nombreux problèmes ouverts ouverts autours de l'interprétation des spectres de masse en protéomique.

# Bibliographie


[ABW+04] R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, et al. UniProt : the universal protein knowledgebase. *Nucleic Acids Research*, 32(Database Issue) :D115–D119, 2004.

[AMB+09] E. Ahrné, A. Masselot, P.A. Binz, M. Müller, and F. Lisacek. A simple workflow to increase MS2 identification rate by subsequent spectral library search. *Proteomics*, 9(6) :1731–1736, 2009.

[AML10] E. Ahrné, M. Müller, and F. Lisacek. Unrestricted identification of modified proteins using MS/MS. *Proteomics*, 10(4) :671–86, 2010.

[BAA94] A.J. Bleasby, D. Akrigg, and T.K. Attwood. OWL-a non-redundant composite protein sequence database. *Nucleic Acids Research*, 22(17) :3574–3577, 1994.

[Bar90] C. Bartels. Fast algorithm for peptide sequencing by mass spectroscopy. *Biomedical & Environmental Mass Spectrometry*, 19(6) :363–368, 1990.

[BCP07] N. Bandeira, K.R. Clauser, and P.A. Pevzner. Shotgun protein sequencing. *Molecular & Cellular Proteomics*, 6(7) :1123, 2007.

[Bel52] R. Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, 38 :716–719, 1952.

[BHL99] P. Berndt, U. Hobohm, and H. Langen. Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints. *Electrophoresis*, 20(18) :3521–3526, 1999.

[BKLL08] S. Bringans, T.S. Kendrick, J. Lui, and R. Lipscombe. A comparative study of the accuracy of several *de novo* sequencing software packages for datasets derived by matrix-assisted laser desorption/ionisation and electrospray. *Rapid Communications in Mass Spectrometry*, 22(21) :3450–3454, 2008.

[CAB+04] S. Carr, R. Aebersold, M. Baldwin, A.L. Burlingame, K.R. Clauser, and A.I. Nesvizhskii. The need for guidelines in publication of peptide and protein identification data. *Molecular & Cellular Proteomics*, 3(6) :531, 2004.

[CB04] R. Craig and R.C. Beavis. TANDEM : matching proteins with tandem mass spectra. *Bioinformatics*, 20(9) :1466–1467, 2004.

[CCFB06] R. Craig, J.C. Cortens, D. Fenyo, and R.C. Beavis. Using annotated peptide mass spectrum libraries for protein identification. *Journal of proteome research*, 5(8) :1843–1849, 2006.

[CFRT09] F. Cliquet, G. Fertin, I. Rusu, and D. Tessier. Comparison of spectra in unsequenced species. In *Proceedings of the 4th Brazilian Symposium on Bioinformatics (BSB 2009) : Advances in Bioinformatics and Computational Biology*, volume 5676 of *LNCS*, pages 24–35. Springer, 2009.

[CFRT10] F. Cliquet, G. Fertin, I. Rusu, and D. Tessier. Proper alignment of MS/MS spectra from unsequenced species. In *Proceedings of the International Conference on Bioinformatics and Computational Biology (BIOCOMP'10)*, volume 2, pages 766–772. CSREA Press, 2010.

[CHS+95] K.R. Clauser, S.C. Hall, D.M. Smith, J.W. Webb, L.E. Andrews, H.M. Tran, L.B. Epstein, and A.L. Burlingame. Rapid mass spectrometric peptide sequencing and mass matching for characterization of human melnoma proteins isolated by two-dimentional PAGE. *Proceedings of the National Academy of Sciences*, 92 :5072–5076, 1995.

[CMG+03] J. Colinge, A. Masselot, M. Giron, T. Dessingy, and J. Magnin. OLAV : towards high-throughput tandem mass spectrometry data identification. *Proteomics*, 3(8) :1454–1463, 2003.

[DAC+99] V. Dancik, T.A. Addona, K.R. Clauser, J.E. Vath, and P.A. Pevzner. *De novo* peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 6(3-4) :327–342, 1999.

[DLA08] E.W. Deutsch, H. Lam, and R. Aebersold. PeptideAtlas : a resource for target selection for emerging targeted proteomics workflows. *EMBo reports*, 9(5) :429–434, 2008.

[DSO78] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5(Suppl 3) :345–352, 1978.

[EMY94] J.K. Eng, A.L. McCormack, and J.R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11) :976–989, 1994.

[FAW+95] R.D. Fleischmann, M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223) :496, 1995.

[FdCGB+99] J. Fernandez-de Cossio, J. Gonzalez, L. Betancourt, V. Besada, G. Padron, Y. Shimonishi, and T. Takao. Automated interpretation of high-energy collision-induced dissociation spectra of singly protonated peptides by SeqMS, a software aid for *de novo* sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 12(23) :1867–1878, 1999.

[FdCGS+00] J. Fernandez-de Cossio, J. Gonzalez, Y. Satomi, T. Shima, N. Okumura, V. Besada, L. Betancourt, G. Padron, Y. Shimonishi, and T. Takao. Automated interpretation of low-energy collision-induced dissociation spectra by SeqMS, a software aid for *de novo* sequencing by tandem mass spectrometry. *Electrophoresis*, 21(9) :1694–1699, 2000.

[FFB02] H.I. Field, D. Fenyö, and R.C. Beavis. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics*, 2(1) :36–47, 2002.

[Fit00] W.M. Fitch. Homology : a personal view on some of the problems. *Trends in Genetics*, 16(5) :227–231, 2000.

[FMM+89] J.B. Fenn, M. Mann, C.K. Meng, S.F. Wong, and C.M. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926) :64, 1989.

[FMW+06] B.E. Frewen, G.E. Merrihew, C.C. Wu, W.S. Noble, and M.J. MacCoss. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Analytical Chemistry*, 78(16) :5678–5684, 2006.

[FP05] A.M. Frank and P.A. Pevzner. PepNovo : *de novo* peptide sequencing via probabilistic network modeling. *Analytical chemistry*, 77(4) :964, 2005.

[Fra09a] A.M. Frank. A Ranking-Based Scoring Function for Peptide- Spectrum Matches. *Journal of Proteome Research*, 8(5) :2241–2252, 2009.

[Fra09b] A.M. Frank. Predicting intensity ranks of peptide fragment ions. *Journal of Proteome Research*, 8(5) :2226–2240, 2009.

[FSN+07] A.M. Frank, M.M. Savitski, M.L. Nielsen, R.A. Zubarev, and P.A. Pevzner. *De novo* peptide sequencing and identification with precision mass spectrometry. *Journal of Proteome Research*, 6(1) :114–123, 2007.

[FTP05] A.M. Frank, S. Tanner, and P.A. Pevzner. Peptide sequence tags for fast database search in mass-spectrometry. *Journal of Proteome Research*, 4(4) :1287–1295, 2005.

[GM01] R. Gras and M. Müller. Computational aspects of protein identification by mass spectrometry. *Current Opinion in Molecular Therapeutics*, 3(6) :526, 2001.

[GMG+99] R. Gras, M. Müller, E. Gasteiger, S. Gay, P.A. Binz, W. Bienvenut, C. Hoogland, J.C. Sanchez, A. Bairoch, D.F. Hochstrasser, et al. Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis*, 20(18) :3535–3550, 1999.

[HBS+93] W.J. Henzel, T.M. Billeci, J.T. Stults, S.C. Wong, C. Grimley, and C. Watanabe. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proceedings of the National Academy of Sciences*, 90(11) :5011, 1993.

[Her05] P. Hernandez. *Peptide identification by tandem mass spectrometry : a tag-oriented open-modification search method*. PhD thesis, Université de Genève, Faculté des sciences, 2005.

[HG10] W. Haerty and G.B. Golding. Low-complexity sequences and single amino acid repeats : not just “junk” peptide sequences. *Genome*, 53(10) :753–762, 2010.

[HGFA03] P. Hernandez, R. Gras, J. Frey, and R.D. Appel. Popitam : towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. *Proteomics*, 3(6) :870–878, 2003.

[HHMM10] E.J. Hsieh, M.R. Hoopmann, B. MacLean, and M.J. MacCoss. Comparison of database search strategies for high precursor mass accuracy MS/MS data. *Journal of Proteome Research*, 9(2) :1138–43, 2010.

[HHS03] M. Havilio, Y. Haddad, and Z. Smilansky. Intensity-based statistical scorer for tandem mass spectrometry. *Analytical Chemistry*, 75(3) :435–44, 2003.

[HLCJJ04] A. Heredia-Langner, W.R. Cannon, K.D. Jarman, and K.H. Jarman. Sequence optimization as an alternative to de novo analysis of tandem mass spectrometry data. *Bioinformatics*, 20(14) :2296, 2004.

[HWS03] W.J. Henzel, C. Watanabe, and J.T. Stults. Protein identification : the origins of peptide mass fingerprinting. *Journal of the American Society for Mass Spectrometry*, 14(9) :931–942, 2003.

[JBA+10] L. Ji, T. Barrett, O. Ayanbule, D.B. Troup, D. Rudnev, R.N. Muertter, M. Tomashevsky, A. Soboleva, and D.J. Slotta. NCBI peptidome : a new repository for mass spectrometry proteomics data. *Nucleic Acids Research*, 38(Database issue) :D731, 2010.

[JLW+08] S.Y. Jung, Y. Li, Y. Wang, Y. Chen, Y. Zhao, and J. Qin. Complications in the assignment of 14 and 28 da mass shift detected by mass spectrometry as in vivo methylation from endogenous proteins. *Analytical Chemistry*, 80(5) :1721–1729, 2008.

[JQCG93] P. James, M. Quadroni, E. Carafoli, and G. Gonnet. Protein identification by mass profile fingerprinting. *Biochemical and Biophysical Research Communications*, 195(1) :58–64, 1993.

[JRP06] S. Jackson, S. Rounsley, and M. Purugganan. Comparative sequencing of plant genomes : choices to make. *The Plant Cell Online*, 18(5) :1100, 2006.

[JT02] R.S. Johnson and J.A. Taylor. Searching sequence databases via *de novo* peptide sequencing by tandem mass spectrometry. *Molecular Biotechnology*, 22(3) :301–315, 2002.

[KBH85] M. Karas, D. Bachmann, and F. Hillenkamp. Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. *Analytical Chemistry*, 57(14) :2935–2939, 1985.

[KEJ+08] J. Klimek, J.S. Eddes, S. Jackson, A. Peterson, S. Letarte, P.R. Gafken, J.E. Katz, P. Mallick, H. Lee, A. Schmidt, R. Ossala, J.K. Eng, R. Aebersold, and D.B. Martin. The standard protein mix database : a diverse data set to assist in the production of improved peptide and protein identification software tools. *Journal of Proteome Research*, 7(1) :96–103, 2008.

[KEZ+05] A. Keller, J. Eng, N. Zhang, X. Li, and R. Aebersold. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Molecular Systems Biology*, 1(1), 2005.

[LBB+07] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21) :2947, 2007.

[LC03] B. Lu and T. Chen. A suffix tree approach to the interpretation of tandem mass spectra : applications to peptides of non-specific digestion and post-translational modifications. *Bioinformatics*, 19(Suppl 2) :113–121, 2003.

[LDE+07] H. Lam, E.W. Deutsch, J.S. Eddes, J.K. Eng, N. King, S.E. Stein, and R. Aebersold. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*, 7(5) :655–667, 2007.

[LPB+10] C. Larré, S. Penninck, B. Bouchet, V. Lollier, O. Tranquet, S. Denery-Papini, F. Guillon, and H. Rogniaux. *Brachypodium distachyon* grain : identification and subcellular localization of storage proteins. *Journal of Experimental Botany*, 61(6) :1771–1783, 2010.

[M+98] K. Mitchelhill et al. Delta Mass : a database of protein post translational modifications. http://www.abrf.org/index.cfm/dm.home, 1998.

[Mat] Matrix Science. http ://www.matrixscience.com/distiller.html.

[Mat07a] R. Matthiesen. Methods, algorithms and tools in computational proteomics : a practical point of view. *Proteomics*, 7(16) :2815–2832, 2007.

[Mat07b] R. Matthiesen. Virtual Expert Mass Spectrometrist v3. 0 : an integrated tool for proteome analysis. *Methods in Molecular Biology (Clifton, NJ)*, 367 :121, 2007.

[MBS+04] R. Matthiesen, J. Bunkenborg, A. Stensballe, O.N. Jensen, K.G. Welinder, and G. Bauw. Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2. 0. *Proteomics*, 4(9) :2583–2593, 2004.

[MHR93] M. Mann, P. Højrup, and P. Roepstorff. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biological Mass Spectrometry*, 22(6) :338, 1993.

[MLWB03] R. Matthiesen, M. Lundsgaard, K.G. Welinder, and G. Bauw. Interpreting peptide mass spectra by VEMS. *Bioinformatics*, 19(6) :792, 2003.

[MTH+05] R. Matthiesen, M.B. Trelle, P. Hojrup, J. Bunkenborg, and O.N. Jensen. VEMS 3.0 : algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. *Journal of Proteome Research*, 4(6) :2338–2347, 2005.

[MW94] M. Mann and M. Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, 66(24) :4390–4399, 1994.

[MZH+03] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie. PEAKS : powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17(20) :2337–2342, 2003.

[NA05] A.I. Nesvizhskii and R. Aebersold. Interpretation of shotgun proteomic data : The protein inference problem. *Molecular & Cellular Proteomics*, 4(10) :1419–1440, 2005.

[NW70] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3) :443–453, 1970.

[PDT00] P.A. Pevzner, V. Dancik, and C.L. Tang. Mutation-tolerant protein identification by mass spectrometry. *Journal of Computational Biology*, 7(6) :777–787, 2000.

[PEH+04] P.G.A. Pedrioli, J.K. Eng, R. Hubley, M. Vogelzang, E.W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R.H. Angeletti, R. Apweiler, et al. A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnology*, 22(11) :1459–1466, 2004.

[PFM+06] S. Pevtsov, I. Fedulova, H. Mirzaei, C. Buck, and X. Zhang. Performance evaluation of existing *de novo* sequencing algorithms. *Journal of Proteome Research*, 5(11) :3018–28, 2006.

[PHB93] D.J.C. Pappin, P. Hojrup, and A.J. Bleasby. Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology*, 3(6) :327–332, 1993.

[PMC07] E. Pitzer, A. Masselot, and J. Colinge. Assessing peptide *de novo* sequencing algorithms performance on large and diverse data sets. *Proteomics*, 7(17) :3051–4, 2007.

[PMDT01] P.A. Pevzner, Z. Mulyukov, V. Dancik, and C.L. Tang. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Research*, 11(2) :290, 2001.

[PP33] A. Payen and J.F. Persoz. Mémoire sur la diastase, les principaux produits de ses réactions, et leurs applications aux arts industriels. In *Annales de Chimie et de Physique*, volume 53, pages 73–92, 1833.

[PPCC99] D.N. Perkins, D.J.C. Pappin, D.M. Creasy, and J.S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18) :3551–3567, 1999.

[PRM+02] M. Palmblad, M. Ramström, K.E. Markides, P. Håkansson, and J. Bergquist. Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry. *Analytical Chemistry*, 74(22) :5826–5830, 2002.

[RF84] P. Roepstorff and J. Fohlman. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical Mass Spectrometry*, 11(11), 1984.

[RKM+09] B.R. Renard, M. Kirchner, F. Monigatti, A.R. Ivanov, J Rappsilber, D. Winter, J.A.J. Steen, F.A. Hamprecht, and H. Steen. When less can yield more - Computational preprocessing of MS/MS spectra for peptide identification. *Proteomics*, 9(21) :4978–4984, 2009.

[SB09] D. Sumpton and W. Bienvenut. Coomassie stains : are they really mass spectrometry compatible ? *Rapid Communications in Mass Spectrometry*, 23(10) :1525–1529, 2009.

[SBE09] D.J. Slotta, T. Barrett, and R. Edgar. NCBI peptidome : a new public repository for mass spectrometry peptide identification. *Nature Biotechnology*, 27(7) :600–601, 2009.

[SDT+04] B.C. Searle, S. Dasari, M. Turner, A.P. Reddy, D. Choi, P.A. Wilmarth, A.L. McCormack, L.L. David, and S.R. Nagalla. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. *Analytical Chemistry*, 76(8) :2220–2230, 2004.

[SDW+05] B.C. Searle, S. Dasari, P.A. Wilmarth, M. Turner, A.P. Reddy, L.L. David, and S.R. Nagalla. Identification of protein modifications using MS/MS de novo sequencing and the OpenSea alignment algorithm. *Journal of Proteome Research*, 4(2) :546, 2005.

[Spe04] B. Spengler. *De novo* sequencing, peptide composition analysis, and composition-based sequencing : a new strategy employing accurate mass determination by Fourier transform ion cyclotron resonance mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 15(5) :703–714, 2004.

[Sta79] R. Staden. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, 6(7) :2601, 1979.

[TJ98] J.A. Taylor and R.S. Johnson. Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 11(9) :1067–1075, 1998.

[TJ01] J.A. Taylor and R.S. Johnson. Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Analytical Chemistry*, 73(11) :2594–2604, 2001.

[TSF+05] S. Tanner, H. Shu, A. Frank, L.C. Wang, E. Zandi, M. Mumby, P.A. Pevzner, and V. Bafnas. InsPecT : identification of posttranslationally modified peptides from tandem mass spectra. *Analytical Chemistry*, 77(14) :4626–4639, 2005.

[TSYI03] D.L. Tabb, A. Saraf, and J.R. Yates III. GutenTag : high-throughput sequence tagging via an empirically derived fragmentation model. *Analytical Chemistry*, 75(23) :6415–6421, 2003.

[TTZ+05] D. Tsur, S. Tanner, E. Zandi, V. Bafna, and P.A. Pevzner. Identification of post-translational modifications by blind search of mass spectra. *Nature Biotechnology*, 23(12) :1562–1567, 2005.

[TWI+88] K. Tanaka, H. Waki, Y. Ido, S. Akita, Y. Yoshida, T. Yoshida, et al. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, 2(8) :151–153, 1988.

[VCR+09] J.A. Vizcaíno, R. Côté, F. Reisinger, J.M. Foster, J. Rameseder, H. Hermjakob, and L. Martens. A guide to the Proteomics Identification Database proteomics data repository. *Proteomics*, 9(18) :4276–4283, 2009.

[VGM+10] J.P. Vogel, D.F. Garvin, T.C. Mockler, J. Schmutz, D. Rokhsar, M.W. Bevan, K. Barry, M. Harmon-Smith, K. Lail, et al. Genome sequencing and analysis of the model grass brachypodium distachyon. *Nature*, 463(7282) :763–768, 2010.

[VST03] S. Veerassamy, A. Smith, and E.R.M. Tillier. A transition probability model for amino acid substitutions from blocks. *Journal of Computational Biology*, 10(6) :997–1010, 2003.

[WF96] J.C. Wootton and S. Federhen. Analysis of compositionally biased regions in sequence databases. *Mehods in Enzymology*, 266 :554–571, 1996.

[WGB+99] M.R. Wilkins, E. Gasteiger, A. Bairoch, J.C. Sanchez, K.L. Williams, R.D. Appel, and D.F. Hochstrasser. Protein identification and analysis tools in the ExPASy server. *Methods in Molecular Biology (Clifton, NJ)*, 112 :531–552, 1999.

[YIMG+98] J.R. Yates III, S.F. Morgan, C.L. Gatlin, P.R. Griffin, and J.K. Eng. Method to compare collision-induced dissociation spectra of peptides : potential for library searching and subtractive analysis. *Analytical Chemistry*, 70(17) :3557–3565, 1998.

[YPO+05] B. Yan, C. Pan, V.N. Olman, R.L. Hettich, and Y. Xu. A graph-theoretic approach for the separation of b and y ions in tandem mass spectra. *Bioinformatics*, 21(5) :563, 2005.

[YSGH93] J.R. Yates, S. Speicher, P.R. Griffin, and T. Hunkapiller. Peptide mass maps : a highly informative approach to protein identification. *Analytical Biochemistry*, 214(2) :397–408, 1993.

[ZC00] W. Zhang and B.T. Chait. ProFound : an expert system for protein identification using mass spectrometric peptide mapping information. *Analytical Chemistry*, 72(11) :2482–2489, 2000.

# Liste des tableaux


2.1 Dénomination et représentation de la structure des 20 acides aminés. . . . . 9
2.2 Abréviations usuelles et masses des 20 acides aminés. . . . . 10

3.1 Liste des plantes séquencées à ce jour. . . . . 35

4.1 Liste des fragments conservés dans le modèle de paquet dédié aux appareils QTOF. 51

5.1 Liste des protéines contenues dans le mélange analysé pour produire les données de l'ISB. . . . . 62
5.2 Description du jeu de données **spectres_ISB**. . . . . 63
5.3 Nombre de spectres présents dans chacun des jeux de spectres issus de l'analyse de Bradi1g13040.1. . . . . 64
5.4 Taille des banques utilisées pour comparer les spectres spectres_ISB . . . . . 65
5.5 Taille des banques Bradi et Bradi_1g. . . . . 65

6.1 Comparaison de l'AUC de la méthode PacketSpectralAlignment selon l'utilisation d'un score de type SPC ou d'un score prenant en considération les probabilités d'apparition des différents ions (SPC pondéré), sur le jeu de données spectres_ISB et la banque 18mix_PAM40. . . . . 74
6.2 Liste des fragments du modèle paquet avec pour chacun d'eux la pondération utilisée dans le score. . . . . 74
6.3 Comparaison des temps d'exécution de PacketSpectralAlignment et de SpectralAlignment sur le jeu de données spectres_ISB. . . . . 93
6.4 Nombre de peptides identifiés lors de la comparaison des spectres de Brachypodium (**bradi_D1** à **bradi_D8**) sur la banque **Bradi_1g** en utilisant dans un premier temps PacketSpectralAlignment sans l'option de score ajusté, puis dans un second temps avec le score ajusté. . . . . 95
6.5 Comparaison des scores des protéines sur les données **bradi_D1** à **bradi_D8**. . . 100

# Liste des figures


2.1 Représentation 3D d'une protéine d'albumine de sérum bovin (BSA). . . . . 6
2.2 Principe de la transcription de l'ADN en ARNm puis de la traduction de ce dernier en protéine. . . . . 7
2.3 Évolution du nombre d'entrées dans UniProt/TrEMBL au cours des 15 dernières années. . . . . 13
2.4 Démarche classique en protéomique. . . . . 15
2.5 Gel d'électrophorèse 2D. . . . . 16
2.6 Représentation d'un analyseur de masse type *Quadripôle*. . . . . 18
2.7 Spectre MS obtenu après l'analyse d'une protéine avec un appareil de type ESI-TOF. 20
2.8 Principe général d'un spectromètre de masse MS/MS. . . . . 22
2.9 Les différents sites de fragmentation à l'intérieur d'un peptide de quatre acides aminés. . . . . 22
2.10 Deux spectres MS/MS illustrant la variabilité de l'analyse d'un même peptide sur un même appareil. . . . . 23
2.11 Deux spectres MS/MS illustrant l'inégalité de répartition de l'information dans un spectre. . . . . 23
2.12 Spectre MS/MS illustrant la variabilité de l'intensité dans un spectre. . . . . 24

3.1 Interprétation *de novo* d'un spectre. . . . . 26
3.2 Exemple de graphe spectral. . . . . 28
3.3 Principe général de la comparaison de spectres (ou PFF). . . . . 30
3.4 (1.) Exemple de spectre théorique, que nous avons généré, contenant toute l'information utile à l'identification. (2.) Un spectre expérimental issu de l'analyse d'une protéine de Brachypodium à l'aide d'un spectromètre MS/MS de type QTOF. 31
3.5 Illustration de la méthode du produit scalaire pour la comparaison de spectres. Cette figure présente la construction du vecteur représentant chacun des spectres, ainsi que la manière dont se calcule le produit scalaire. *(Source : P. Hernandez)* . 33
3.6 Influence des modifications au sein d'un spectre. . . . . 37
3.7 Augmentation de la taille de la banque de protéines à cause des modifications variables . . . . . 38
3.8 Sélection d'un candidat à l'aide d'un tag. . . . . 39
3.9 Matrice d'aligment de SpectralAlignment. . . . . 40
3.10 Comportement de SpectralAlignment en fonction du nombre de modifications présentes sur un jeu de données générées. . . . . 41
3.11 Impact des modifications sur l'alignement de plusieurs spectres. . . . . 43

4.1 Les différents sites de fragmentation à l'intérieur d'un peptide de quatre acides aminés. . . . . 46
4.2 Conséquences de la symétrie dans un spectre MS/MS. . . . . 47
4.3 Représentation de la notion de paquets dans un spectre MS/MS. . . . . 48

4.4 Spectre MS/MS après introduction d'une modification post-traductionnelle sur le second acide aminé. Les traits continus représentent les ions N-terminaux tandis que les pointillés représentent les ions C-terminaux. Chaque paquet est distingué à l'aide d'une couleur différente. . . . . 49
4.5 En a. un spectre théorique représentant le peptide GLMPRG. En b. le spectre théorique symétrique du même peptide. Les couleurs utilisées permettent de distinguer les différents paquets au sein des spectres. . . . . 50
4.6 Modèle de paquet que nous avons développé pour les appareils de type QTOF. . . 50
4.7 Création d'un spectre expérimental symétrique. Les pics au-dessus de l'axe horizontal sont les pics d'origine du spectre expérimental, les pics sous l'axe horizontal sont les pics complémentaires ajoutés. . . . . 52

5.1 Brachypodium distachyon. *Source : Dawson, J.E. and Hatch, S.L.*. . . . . 64
5.2 Mesure de l'identité entre les séquences peptidiques non modifiées avec leur version modifiée en utilisant différentes matrices PAM. Les mesures ont été réalisées sur la banque 18mix_rice1700. . . . . 66
5.3 Exemple de courbes ROC. *(Source :www.medhyg.ch)* . . . . . 68

6.1 Un spectre MS/MS bruité sur la gauche et avec peu de pics sur la droite. Le rectangle bleu représente une fenêtre de filtrage de largeur égale à 110 daltons. . 70
6.2 Évaluation du comportement de l'algorithme PacketSpectralAlignment -AUC et temps d'exécution- en fonction de la largeur de la fenêtre et du nombre de pics conservés dans la fenêtre -4, 6 ou 8 pics- avec le jeu de données spectres_ISB et la banque 18mix_PAM40. . . . . 72
6.3 Évaluation de l'efficacité d'un filtrage par fenêtre en fonction du nombre de pics conservés dans une fenêtre constante de largeur 110 daltons, sur le jeu de données spectres_ISB et la banque 18mix_PAM40. . . . . 73
6.4 Évaluation de différentes méthodes de score, prenant en compte l'intensité des pics, pour l'alignement d'un paquet en utilisant le jeu de données spectres_ISB et la banque 18mix_PAM40. . . . . 75
6.5 Différents exemples de scores pour différents alignements de paquets. Ne sont représentés ici que les paquets, avec en rouge les pics alignés et en noir les pics non alignés. . . . . 76
6.6 Évaluation de l'impact du filtrage des positions possibles en terme de temps (courbe grise) et en terme de qualité (courbe noire) sur le jeu de données spectres_ISB et la banque 18mix_PAM40. . . . . 78
6.7 Évaluation du filtrage des positions possibles à l'aide d'une fenêtre en terme de temps (courbe grise) et en terme de qualité (courbe noire) sur le jeu de données spectres_ISB et la banque 18mix_PAM40. . . . . 79
6.8 Évaluation de la qualité de la méthode de comparaison pour un filtrage de la banque autorisant au maximum $10\%$ à $30\%$ de différence de masse entre le spectre expérimental et les candidats de la banque. L'AUC lorsque aucun filtrage n'est effectué est aussi donnée à titre indicatif. Cette évaluation a été effectuée sur le jeu de données spectres_ISB et la banque 18MIC_PAM40. . . . . 81

6.9 Comparaison de différentes valeurs de modifications tolérées pour rendre $K$ dépendant de la longueur du peptide, avec le jeu de données spectres_ISB et la banque 18mix_PAM40. . . . . 82
6.10 L'enchaînement des composants de la plate-forme SIFpackets. . . . . 85
6.11 L'enchaînement des composants de la plate-forme SIFpackets avec l'ajout de la variante avec liste d'exclusion. . . . . 90
6.12 Évaluation de la qualité (AUC) des résultats des méthodes PacketSpectralAlignment (PSA) et SpectralAlignment (SA) en fonction du niveau de modification de la banque. . . . . 93
6.13 Courbes ROC représentant les résultats de PacketSpectralAlignement (en noir) et de SpectralAlignment (en gris). Les deux figures représentent les mêmes données, mais la figure du bas est tracée en échelle logarithmique, ce qui permet de mettre en évidence la différence de comportement pour les faibles taux de faux positifs. . 94
6.14 Alignement d'un spectre expérimental comportant une modification identifiée et localisée à l'aide de la plate-forme SIFpackets -spectre du haut- avec un spectre théorique -spectre du bas-. . . . . 96

A.1 Comportement de SpectralAlignment en fonction du nombre de modifications présentes sur un jeu de données générées. . . . . 132

# Liste des exemples

2.1 Substitution d'un acide aminé dans un peptide . . . . . 11
2.2 Insertion d'un acide aminé dans un peptide . . . . . 11
2.3 Suppression d'un acide aminé dans un peptide . . . . . 11
2.4 Comparaison de protéines . . . . . 11
2.5 Séquence de la protéine ACTA_BOVIN au format FASTA . . . . . 13

# List of Algorithms


1 PacketSpectralAlignment : alignement . . . . . 57
2 PacketSpectralAlignment : backtrack . . . . . 59
3 Algorithme de création et de filtrage de la liste des positions possibles . . . . . 77
4 PacketSpectralAlignment avec liste d'exclusion . . . . . 87
5 PacketSpectralAlignment avec liste d'exclusion, suite . . . . . 88
6 Algorithme de calcul du score ajusté . . . . . 91
7 Algorithme de calcul du score d'un peptide lors de la remontée à la protéine. . . . 98

# Table des matières


**1 Introduction** **1**

**2 Notions de biologie et de protéomique** **5**
2.1 Introduction . . . 5
2.2 Des gènes aux protéines . . . 6
2.2.1 La cellule et ses protéines . . . 6
2.2.2 Modifications . . . 10
2.2.3 Banques de protéines . . . 12
2.3 Protéomique et spectrométrie de masse . . . 14
2.3.1 Introduction à l'identification de protéines . . . 14
2.3.2 Séparation des protéines . . . 14
2.3.3 Hydrolyse des protéines . . . 17
2.3.4 Spectrométrie de masse . . . 17

**3 L'identification de protéines en MS/MS - État de l'art et problématique** **25**
3.1 Introduction . . . 25
3.2 L'interprétation *de novo* d'un spectre MS/MS . . . 25
3.2.1 L'interprétation manuelle d'un spectre MS/MS . . . 26
3.2.2 L'interprétation automatisée d'un spectre MS/MS . . . 26
3.2.3 Comparaison des méthodes *de novo* . . . 28
3.3 L'identification par comparaison avec des protéines connues . . . 29
3.3.1 Le principe général de la comparaison de spectres . . . 29
3.3.2 Filtres sur la sélection des peptides théoriques . . . 30
3.3.3 La construction de spectres théoriques . . . 31
3.3.4 Évaluation de la similarité entre deux spectres . . . 32
3.3.5 Bibliothèques Spectrales . . . 33
3.4 Comparaison des approches *de novo* et de PFF . . . 34
3.5 La problématique des modifications sans a priori . . . 34
3.5.1 La protéomique des organismes non séquencés . . . 34
3.5.2 Différents types de modifications des protéines . . . 35
3.5.3 Conséquences d'une modification dans un spectre . . . 36
3.5.4 Comparaison des différentes approches d'identification en présence de modifications . . . 36

**4 PacketSpectralAlignment, une nouvelle méthode de comparaison de spectres** **45**
4.1 Introduction . . . 45
4.2 Notations . . . 45
4.3 Deux notions importantes : Symétrie et Paquets . . . 47
4.3.1 La Symétrie interne au spectre . . . 47
4.3.2 Les Paquets . . . 48

4.4 Modification des spectres . . . 48
4.4.1 Modifications des spectres théoriques . . . 49
4.4.2 Modifications des spectres expérimentaux . . . 51
4.4.3 Particularités de l'alignement des spectres symétriques . . . 52
4.5 Algorithme d'alignement de deux spectres . . . 53
4.5.1 Principe de la programmation dynamique . . . 53
4.5.2 Les paramètres de l'algorithme d'alignement . . . 54
4.5.3 Le résultat d'un alignement . . . 54
4.5.4 Le score mesurant la similarité entre deux spectres . . . 55
4.5.5 L'algorithme d'alignement PacketSpectralAlignment . . . 55

**5 Jeux de données et critères d'évaluation 61**
5.1 Introduction . . . 61
5.2 Jeux de données . . . 61
5.2.1 Jeux de données de spectres . . . 62
5.2.2 Banques de données . . . 65
5.3 Critères d'évaluation . . . 67
5.3.1 Temps d'exécution . . . 67
5.3.2 Qualité des résultats . . . 67

**6 SIFpackets : mettre PacketSpectralAlignment en situation réelle 69**
6.1 Introduction . . . 69
6.2 Amélioration de l'identification des peptides : paramétrage et prétraitements . . . 69
6.2.1 Filtrage des spectres expérimentaux . . . 69
6.2.2 Modèle de score pour l'alignement d'un paquet . . . 72
6.2.3 Filtrage des positions possibles . . . 75
6.2.4 Filtrage de la banque . . . 80
6.2.5 Nombre de modifications tolérées . . . 81
6.2.6 Pénalités de modification . . . 83
6.3 SIFpackets : une plate-forme complète associant spectres et peptides . . . 84
6.3.1 Description de la plate-forme SIFpackets . . . 84
6.3.2 Variante permettant une meilleure prise en compte des modifications . . . 84
6.3.3 Résultats expérimentaux : ISB Dataset . . . 92
6.3.4 Résultats expérimentaux : Brachypodium . . . 93
6.4 Remontée à la protéine . . . 96
6.4.1 Établir un score pour les protéines . . . 97
6.4.2 Résultat sur les données de Brachypodium . . . 99

**7 Conclusions et perspectives 101**
7.1 Conclusions . . . 101
7.2 Perspectives . . . 102

**Bibliographie 105**

**Liste des tableaux 113**

**Liste des figures** **115**

**Liste des exemples** **119**

**Table des matières** **123**

**Index** **125**

**A Évaluation du comportement de SpectralAlignment en présence de modifications** **131**

# Index


acide aminé, 8

banque de protéines, 12
    Brachypodium, 65
    ISB, 65
    UniProt, 12
bibliothèque spectrale, 33

candidat, 31
charge, 47
chromatographie, 15
    en phase liquide, 15
comparaison de spectres, 29
complémentaire
    séquences, 46

*de novo*, 25
    interprétation automatisée, 26
    interprétation manuelle, 26
    pseudo PFF, 27
    reconstruction iterative, 27

électrophorèse, 16
enzyme, 17
évaluation, 67
    qualité, 67
    temps, 67

filtrage
    positions possibles, 75
    spectres expérimentaux, 69
fragmentation, 21, 46

graphe spectral, 28

hydrolyse, 17

insertion, 11
ion, 46
ionisation, 17

jeu de données, 61
    Brachypodium, 63
    ISB, 62

masse
    d'un ion, 46
    d'une séquence d'acide aminés, 46
mélange de protéines, 21
modification, 10, 11, 35
    ajout de, 65
    post-traductionnelle, 11
MS, 19
MS/MS, 21
mutation, 10

PacketSpectralAlignment, 55
PAM, 66
paquet, 48
    modèle de, 51
    point de référence, 49
peptide, 8
    candidat, 31
PFF, voir comparaison de spectres
pic
    C-terminal, 47
    N-terminal, 47
PMF, 20
point de référence, 49
positions possibles, 53, 75
précurseur, 23
prétraitement, 19
programmation dynamique, 53
protéine, 6
    structure, 8
    synthèse, 7
protéomique, 14

score, 32, 55, 72
    corrélation croisée, 32
    produit scalaire, 32
    SPC, 32

shotgun, voir mélange de protéines
similarité, 32, 55
SpectralAlignment, 40
spectre
    expérimental symétrique, 51
    MS, 19
    MS/MS, 22
    observable, 33
    théorique, 31
    théorique symétrique, 49
spectrométrie de masse, 17
substitution, 11
suppression, 11
symétrie
    interne au spectre, 47

tag, 8
trypsine, 17

# Annexes

# ANNEXE A

# Évaluation du comportement de SpectralAlignment en présence de modifications

Nous avons évalué le comportement de SpectralAlignment (SA) sur un ensemble de données simulées. Nous avons tout d'abord généré un ensemble de 1000 peptides aléatoires de taille comprise dans $[10; 25]$ afin de constituer une banque qui sert à créer les spectres théoriques. Chaque peptide de cette banque est modifié en 5 versions différentes en appliquant de 0 à 4 substitutions aléatoires d'acides aminés. Ces peptides modifiés sont utilisés pour créer 5 ensembles de spectres expérimentaux (un ensemble par nombre de modifications appliquées). Les spectres expérimentaux sont constitués en utilisant les 9 pics les plus fréquemment rencontrés, pics dont la présence dans le spectre est décidée aléatoirement en fonction de leur probabilité d'apparition [DAC$^+$99]. Ensuite, du bruit est introduit dans ces spectres, en ajoutant $50\%$ de pics supplémentaires à des masses choisies aléatoirement. Tous les tests ont été effectués à une précision de 1 dalton. Pour chaque spectre expérimental $S_e$, nous appelons peptide cible de $S_e$ (noté TP($S_e$)) la séquence du peptide d'origine de notre banque, celle qui a été modifiée pour produire $S_e$.

Chaque spectre théorique est comparé à chaque spectre expérimental. Les scores (ici, le nombre de pics communs du meilleur alignement) résultant de chaque comparaison sont mémorisés et utilisés pour trier les ensembles de peptides pour chacun des $S_e$. Dans le cas idéal, pour un spectre expérimental $S_e$, TP($S_e$) doit obtenir le meilleur score et donc avoir pour rang 1. Il est donc possible, en observant le rang des peptides cibles, d'évaluer la capacité de SpectralAlignement à prendre en compte des modifications. C'est pourquoi nous utilisons le rang moyen du peptide cible ainsi que la proportion de peptide cible ayant obtenu le rang 1, comme indicateurs afin d'évaluer l'algorithme, comme cela est représenté dans la Figure A.1.

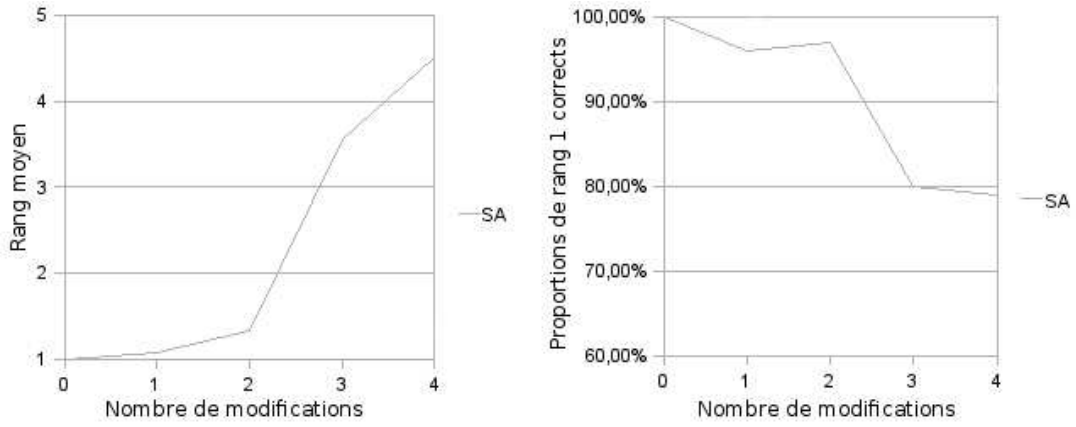


Figure A.1 – Comportement de SpectralAlignment en fonction du nombre de modifications présentes sur un jeu de 1000 spectres créés *in silico*. Le rang représente ici le classement de l'identification correcte parmi tous les peptides candidats ordonnés par score décroissant. Il est visible ici qu'avec plus de deux modifications, le taux d'identification décroît considérablement.

# Des spectres MS/MS à l'identification des protéines - Interprétation des données issues de l'analyse d'un mélange de protéines d'un organisme non séquencé

Freddy CLIQUET

**Résumé**

La spectrométrie de masse est une technique utilisée en protéomique pour identifier des protéines inconnues dans un échantillon. Le spectromètre mesure la masse de fragments de la protéine et fournit ainsi des spectres expérimentaux qui sont des représentations, sous forme de séries de pics, de la présence de ces différents fragments. En étudiant ces spectres, nous espérons pouvoir identifier la protéine d'origine en la retrouvant dans une banque.
L'objectif de cette thèse est de proposer de nouvelles méthodes permettant d'étudier ces spectres. Cependant, ces méthodes doivent fonctionner sur des organismes non séquencés. Dans ce cas particulier, nous ne retrouverons pas exactement ces protéines dans la banque, mais uniquement des protéines qui y ressemblent.

Nous proposons tout d'abord un nouvel algorithme dit de comparaison de spectres : PacketSpectralAlignment. Cet algorithme permet de comparer des spectres expérimentaux à des spectres créés à partir des données contenues dans la banque, et ce, même en présence de modifications. Cette comparaison permet l'association de chacun des spectres à un peptide de cette banque. Ensuite, nous détaillerons différents prétraitements et filtrages permettant d'améliorer l'exploitation de notre nouvel algorithme. Tous ces éléments sont intégrés dans une plate-forme intitulé SIFpackets. Enfin, nous validons les résultats de PacketSpectralAlignment ainsi que de SIFpackets sur différents jeux de données réelles.

**Mots-clés :** Protéomique, Spectrométrie de masse, Comparaison de spectres, Identification de protéines, Modifications post-traductionnelles, Algorithmes, Programmation dynamique

**Abstract**

Mass spectrometry is a general method used in proteomics to identify unknown proteins in a sample. The mass spectrometer measures the masses of several protein's fragments and provide spectra. A spectrum is a series of peaks that indicate the presence of those fragments. By studying these spectra, we aim at retrieving the analyzed protein in a reference databank. In this thesis, we propose a new method to study these spectra. However, our solution must be able to work on proteins coming from unsequenced species, which means that we can't find exactly the same proteins in the databank, only similar ones.

At first, we propose a new spectra comparison algorithm: PacketSpectralAlignment. This algorithm allows to compare experimental spectra produced by a mass spectrometer to spectra created from the reference databank data in presence of modifications. This comparison allows to associate to each spectrum, a peptide from the databank.
Then, we explain several preprocessing and filtering methods that enhance the results of our new algorithm. All of those methods are used in the SIFpackets framework.
Finally, we validate PacketSpectralAlignment and SIFpackets results using several experimental datasets.

**Keywords:** Proteomics, Mass spectrometry, Spectra comparison, Protein Identification, Post translational modifications, Algorithms, Dynamic programming